DOCUMENT RESUME

ED 342 777 TM 016 513

AUTHOR Cook, Linda L.; And Others
TITLE Equating Achievement Tests Using Samples Matched on Ability. College Board Report No. 90-2.
INSTITUTION College Entrance Examination Board, New York, N.Y.
REPORT NO ETS-RR-90-10
PUB DATE 90
NOTE 66p.; A previous version of this paper was presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989).
AVAILABLE FROM College Board Publications, Box 886, New York, NY 10101 ($7.00).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Academic Ability; *Achievement Tests; Biology; Chemistry; *College Entrance Examinations; Comparative Testing; Data Collection; *Equated Scores; French; Higher Education; Language Tests; Mathematics Tests; Robustness (Statistics); *Sampling; Science Tests; Social Studies; Test Items; United States History
IDENTIFIERS Anchor Tests; *College Board Achievement Tests; Common Item Effect; Equipercentile Equating; Levine Equating Method; *Parallel Test Forms; Tucker Common Item Equating Method

ABSTRACT

The equating of reasonably parallel forms of College Board Achievement Tests in biology, chemistry, mathematics level II, American history and social studies, and French is discussed. Results of the following five equating methods are compared: (1) Tucker; (2) Levine equally reliable; (3) Levine unequally reliable; (4) frequency estimation equipercentile; and (5) chained equipercentile. These methods are used with an internal common-item anchor-test data collection design and three sampling strategies (random samples from populations similar in ability level, random samples from populations of dissimilar ability, and samples from dissimilar populations constructed to be similar by matching on the basis of a covariate such as the distribution of scores on a set of common items). Results indicate that it may be difficult, and in some cases impossible, to equate achievement tests using new-form and old-form samples from populations differing in ability level. All these equating methods appear to be affected by group differences in ability, with the Tucker and frequency estimation equipercentile methods the most affected, and the chained equipercentile and the two Levine procedures the most robust. Matching cannot be recommended for rectifying sample ability-level differences. There are 17 tables of study data, 33 figures, and a list of 16 references. (Author/SLD)

ED342777

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN **OTHER THAN PAPER COPY** HAS BEEN GRANTED BY

ROBERT G. CAMERON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

College Board Report

No. 90-2

Equating Achievement Tests Using Samples Matched on Ability

Linda L. Cook
Daniel R. Eignor
Alicia P. Schmitt

TM016513

BEST COPY AVAILABLE

Equating Achievement Tests Using Samples Matched on Ability

Linda L. Cook
Daniel R. Eignor
Alicia P. Schmitt

College Board Report No. 90-2
ETS RR No. 90-10

College Entrance Examination Board, New York, 1990

Linda L. Cook is at Educational Testing Service, Princeton, New Jersey.

Daniel R. Eignor is at Educational Testing Service, Princeton, New Jersey.

Alicia P. Schmitt is at Educational Testing Service, Princeton, New Jersey.

Acknowledgments

The authors would like to recognize Miriam Feigenbaum, Karen Carroll, Chitra Lele, Robin Hochman, Kristin Wichert, Susan Bryce, and Bernadette McIntosh for their contributions to the data preparation and analysis aspects of this study and Karen Damiano for typing this paper. In addition, the authors would like to thank Martha Stocking, Neal Kingston, William Angoff, and Nancy Petersen for their insightful reviews of earlier drafts of this report.

A previous version of this paper was presented at the annual meeting of the American Educational Research Association, San Francisco, 1989. This study was supported by The College Board through Joint Staff Research and Development Committee Funding.

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in College Board Reports do not necessarily represent official College Board position or policy.

The College Board is a nonprofit membership organization that provides tests and other educational services for students, schools, and colleges. The membership is composed of more than 2,700 colleges, schools, school systems, and education associations. Representatives of the members serve on the Board of Trustees and advisory councils and committees that consider the programs of the College Board and participate in the determination of its policies and activities.

Additional copies of this report may be obtained from College Board Publications, Box 886, New York, New York 10101. The price is $7.

Copyright © 1990 by College Entrance Examination Board. All rights reserved.
College Board, Scholastic Aptitude Test, SAT, and the acorn logo are registered trademarks of the College Entrance Examination Board.
Printed in the United States of America.

CONTENTS

Abstract 1

Introduction 1

Methodology 2
- Tests, Test Forms, and Samples 2
- Conventional Item Statistics (Equated Deltas) 5
- Score Equating Methods 5
- Criterion Measures 5

Results and Discussion 6
- Item Analyses 6
- Equatings 8
- Conclusions 17

References 17

Figures

1. Distributions of scores and percentages below on the 36 common items used in matching the Biology Test spring old-form group to the fall new-form group. 19

2. Biology SDQ question 15 and specifications for the recoding of responses for cross-tabulation and matching purposes; correlations of recoded responses with Biology Achievement Test scores for fall new-form and spring old-form groups. 20

3. Biology SDQ question 58 and specifications for the recoding of responses for cross-tabulation and matching purposes; correlations of recoded responses with Biology Achievement Test scores for fall new-form and spring old-form groups. 21

4. Cross-tabulations of recoded responses to SDQ questions 15 and 58 used in matching the Biology spring old-form group to the fall new-form group. 22

5. Distributions of scores and percentages below on the 19 common items used in matching the Mathematics Level II Test spring old-form group to the fall new-form group. 23

6. Distributions of scores and percentages below on the 20 common items used in matching the American History and Social Studies Test spring old-form group to the fall new-form group. 24

7. Distributions of scores and percentages below on the 20 common items used in matching the Chemistry Test spring old-form group to the fall new-form group. 25

8. Distributions of scores and percentages below on the 21 common items used in matching the French Test spring old-form group to the fall new-form group. 26

9. French SDQ question 8 and specifications for the recoding of responses for cross-tabulation and matching purposes; correlations of recoded responses with French Achievement Test scores for fall new-form and spring old-form groups. 27

10. French SDQ question 14 and specifications for the recoding of responses for cross-tabulation and matching purposes; correlations of recoded responses with French Achievement Test scores for fall new-form and spring old-form groups. 28

11. Cross-tabulations of recoded responses to SDQ questions 8 and 14 used in matching the French spring old-form group to the fall new-form group. 29

12. Question on Background Questionnaire in French used in matching the French spring old-form group to the fall new-form group. 30

13. Distributions of responses to question on Background Questionnaire in French used in matching the French spring old-form group to the fall new-form group. 31

14. Plots of Biology fall new-form versus fall old-form equated deltas for the one fall-fall sample combination and versus spring old-form equated deltas for the three fall-spring sample combinations. 32

15. Plots of Mathematics Level II fall new-form versus fall old-form equated deltas for the one fall-fall sample combination and versus spring old-form equated deltas for the two fall-spring sample combinations. 33

16. Plots of American History and Social Studies fall new-form versus fall old-form equated deltas for the one fall-fall sample combination and versus spring old-form equated deltas for the two fall-spring sample combinations. 34

17. Plots of Chemistry fall new-form versus fall old-form equated deltas for the one fall-fall sample combination and versus spring old-form equated deltas for the two fall-spring sample combinations. 35

18. Plots of French fall new-form versus fall old-form equated deltas for the one fall-fall sample combination and versus spring old-form equated deltas for the four fall-spring sample combinations. 36

19. Biology equating-difference plots (fall-spring random or matched groups equatings minus fall-fall random groups criterion equating). 38

20. Biology equating-difference plots (fall-fall random and fall-spring random or matched groups equatings minus fall-fall random groups Tucker criterion equating). 39

21. Plot of Biology projected new-form scaled-score means for all equating-method and equating-sample combinations. 40

22. Mathematics Level II equating-difference plots (fall-spring random or matched groups equatings minus fall-fall random groups criterion equating). 41

23. Mathematics Level II equating-difference plots (fall-fall random and fall-spring random or matched groups equatings minus fall-fall random groups Tucker criterion equating). 42

24. Plot of Mathematics Level II projected new-form scaled-score means for all equating-method and equating-sample combinations. 43

25. American History and Social Studies equating-difference plots (fall-spring random or matched groups equatings minus fall-fall random groups criterion equating). 44

26. American History and Social Studies equating-difference plots (fall-fall random and fall-spring random or matched groups equatings minus fall-fall random groups Tucker criterion equating). 45

27. Plot of American History and Social Studies projected new-form scaled-score means for all equating-method and equating-sample combinations. 46

28. Chemistry equating-difference plots (fall-spring random or matched groups equatings minus fall-fall random groups criterion equating). 47

29. Chemistry equating-difference plots (fall-fall random and fall-spring random or matched groups equatings minus fall-fall random groups Tucker criterion equating). 48

30. Plot of Chemistry projected new-form scaled-score means for all equating-method and equating-sample combinations. 49

31. French equating-difference plots (fall-spring random or matched groups equatings minus fall-fall random groups criterion equating). 50

32. French equating-difference plots (fall-fall random and fall-spring random or matched groups equatings minus fall-fall random groups Tucker criterion equating). 51

33. Plot of French projected new-form scaled-score means for all equating-method and equating-sample combinations. 52

Tables

1. Numbers of Items in New and Old Forms and in Common-Item Sets for Tests in Study 53

2a.–2e. Content and Skills Specifications for ATP Achievement Tests 53

3. Raw-Score Summary Statistics for Total Tests and the Common-Item Set for Samples Used in Equatings—Biology 55

4. Raw-Score Summary Statistics for Total Tests and the Common-Item Set for Samples Used in Equatings—Mathematics Level II 55

5. Raw-Score Summary Statistics for Total Tests and the Common-Item Set for Samples Used in Equatings—American History and Social Studies 55

6. Raw-Score Summary Statistics for Total Tests and the Common-Item Set for Samples Used in Equatings—Chemistry 55

7. Raw-Score Summary Statistics for Total Tests and the Common-Item Set for Samples Used in Equatings—French 56

8. Correlation Coefficients for Item-Difficulty Estimates (Deltas) and New-Form Total-Test Summary Statistics for Different Sample Combinations—Biology 56

9. Correlation Coefficients for Item-Difficulty Estimates (Deltas) and New-Form Total-Test Summary Statistics for Different Sample Combinations—Mathematics Level II 56

10. Correlation Coefficients for Item-Difficulty Estimates (Deltas) and New-Form Total-Test Summary Statistics for Different Sample Combinations—American History and Social Studies 56

11. Correlation Coefficients for Item-Difficulty Estimates (Deltas) and New-Form Total-Test Summary Statistics for Different Sample Combinations—Chemistry 56

12. Correlation Coefficients for Item-Difficulty Estimates (Deltas) and New-Form Total-Test Summary Statistics for Different Sample Combinations—French 57

13. New-Form Scaled-Score Summary Statistics Resulting from Equating-Method and Equating-Sample Combinations—Biology 57

14. New-Form Scaled-Score Summary Statistics Resulting from Equating-Method and Equating-Sample Combinations—Mathematics Level II 57

15. New-Form Scaled-Score Summary Statistics Resulting from Equating-Method and Equating-Sample Combinations—American History and Social Studies 58

16. New-Form Scaled-Score Summary Statistics Resulting from Equating-Method and Equating-Sample Combinations—Chemistry 58

17. New-Form Scaled-Score Summary Statistics Resulting from Equating-Method and Equating-Sample Combinations—French 58

ABSTRACT

The equating of reasonably parallel forms of College Board Achievement Tests in Biology, Chemistry, Mathematics Level II, American History and Social Studies, and French is discussed in this paper. The results of five equating methods are compared: (1) Tucker, (2) Levine equally reliable, (3) Levine unequally reliable, (4) frequency estimation equipercentile, and (5) chained equipercentile. These methods are used with an internal common-item anchor-test data collection design. Three sampling strategies were evaluated: (1) random samples from populations similar in ability level, (2) random samples from populations dissimilar in ability level, and (3) samples from populations dissimilar in ability level that have been constructed to be similar in ability level by matching on the basis of a covariate, such as the distribution of scores on a set of common items. The criteria for comparison in all cases were the results of the Tucker procedure used with random samples from populations similar in ability level. These results were used as the criterion for equating results because they represent results obtained under the most optimal operational conditions.

The results of the study indicate that it may be difficult, and in some cases impossible, to equate achievement tests using new- and old-form samples obtained from populations that are different in ability level. All equating methods investigated in this study appear to be affected by group differences in ability. The equating methods that appear to be the most affected by these differences are the Tucker and frequency estimation equipercentile procedures. The methods that appear to be the most robust to group differences in ability are the chained equipercentile and the two Levine procedures.

Matching on the basis of observed scores on a set of internal common items does not remedy the situation. In general, matching produces results, particularly scaled-score means, for all equating procedures that are similar but that over- or underestimate the criterion scaled-score means. Because the results (i.e., scaled-score means) are similar across methods, the effect can be quite misleading in that, in the absence of a criterion, one could conclude that because consistent results are obtained across methods the results are close to "truth." This was not found to be the case for the situations investigated in this study, and matching cannot be recommended as a procedure for rectifying the problem of sample ability-level differences.

INTRODUCTION

Practitioners working on large-scale admissions testing programs are typically faced with the situation of having to equate test forms taken by groups that differ systematically in ability. In this situation, the equating is usually performed using an anchor-test design, where the anchor can be a set of common items embedded in the forms to be equated. An additional concern of individuals engaged in the equating of achievement tests, given at multiple administrations that span the school year, is that such tests have been specifically designed to reflect course content. This may affect students taking the tests at different points in their coursework in different ways. These students may not only differ in ability, but they may also not constitute samples from the same population. For instance, students who elect to take a test at a spring administration are usually finishing a course of instruction in the content covered by the test. Students who typically elect to take the test at a fall administration may not have had formal coursework for some time.

Underlying any commonly accepted definition of equated scores is the requirement that the equating transformation be sample independent, that is, the raw-score to scale-score conversion function should be the same, regardless of the sample or samples from the population from which it is derived (see Angoff 1984 or Holland and Rubin 1982). In the context of equating achievement tests, this requirement would suggest, for instance, that the equating transformation for a new test form should be the same, regardless of the time of year the sample data used in the equating experiment were collected.

Achievement Tests administered for the College Board's Admissions Testing Program (ATP) have historically been equated using an anchor-test design, using a set of internal common items as the anchor test (see Angoff 1984 for a description of this type of design). New forms of the Achievement Tests are administered and equated using new- and old-form data collected at an administration occurring in the same month but in different years. Typically, the old-form sample is selected from an administration that occurred perhaps two or three years prior to the administration of the new form. In spite of the time lag between the administrations of the new and old forms, the two samples are usually similar in composition, ability level, and preparation or coursework relevant to the particular achievement test. Recently, it has become necessary, because of an increase in the volume of new and revised forms requiring equating, to consider introducing new forms at administrations where comparable old-form data may not exist. This raises a question regarding the adequacy of an equating transformation that is obtained, for instance, when a new-form sample from a spring administration of a test is used along with an old-form sample from a fall administration. There is existing evidence that this situation may seriously affect equating results.

Cook, Eignor, and Taft (1988) examined the results of equating two forms of the Achievement Test in Biology. In their study, one old-form sample and two different new-form samples provided the data for equating. The old-form sample was randomly selected from a fall administration of the test. One of the new-form samples was randomly selected from a spring administration of the test; the second sample was randomly selected from a fall administration. Groups of students electing to take the Biology Test at the fall and the spring administrations vary greatly in the abilities measured by the test. Students taking the test in the spring typically have finished or are about to finish a course in the content covered by the test, whereas those taking the test in the fall may have taken the relevant biology course 6

to 18 months prior to the test administration. Cook et al. equated the two forms of the Biology Test using conventional linear procedures (Tucker or Levine equating models, depending on ability-level differences; see Angoff 1984) and chained equipercentile (Design V in Angoff 1984) and three parameter logistic (3-PL) model Item Response Theory (IRT) curvilinear equating procedures (see Lord 1980). All equatings were first carried out using the spring new-form and fall old-form pairing and then using the fall new-form and fall old-form pairing. Equating results were quite different for these two sample combinations. The equatings based on the combination of the spring new-form and fall old-form samples resulted in scaled-score means at least 15 points higher than those based on the combination of the fall new-form and fall old-form samples. There was also a good deal of variation across the equating results for the various methods performed using data from the spring new-form and fall old-form samples. The results of the Cook et al. study clearly demonstrated the need to define the population of students for whom the equating transformation is applicable.

Recently, in an attempt to ameliorate differences in anchor-test equating results caused by samples that differ in ability, researchers have begun to study the effects on equating procedures of matching one of the samples to the other, usually through scores on either a set of internal common items or an external anchor test. Lawrence and Dorans (1988) studied the effects of matching, using anchor-test scores, on Scholastic Aptitude Test (SAT) verbal and mathematical equatings. The results of the Lawrence and Dorans study suggested that matching the differing ability samples in the Cook, Eignor, and Taft (1988) study (i.e., the spring new-form and fall old-form samples) on an available covariate prior to equating might diminish the differences in equating results seen in that study. In addition, the equatings of the combination of the fall new-form and fall old-form samples from the Cook, Eignor, and Taft study provided a useful evaluative criterion for any matching done with the spring new-form and fall old-form combination in that the combination of the two fall samples provided equating results that were obtained under what has traditionally been considered to be the most satisfactory operational conditions.

With the above in mind, Cook, Eignor, and Schmitt (1988) examined the effects on IRT and conventional (Tucker, Levine unequally reliable, and chained equipercentile) equating results of matching fall-spring new- and old-form Biology samples, using two different covariate measures: (1) observed scores on the internal common-item equating block, and (2) responses to selected questions on the Student Descriptive Questionnaire (SDQ), a questionnaire that examinees respond to on a voluntary basis when filling out their registration forms for the test. An older version of the SDQ was used in the study; it has since been revised. The criteria used in the Cook, Eignor, and Schmitt study were the Tucker equatings based on the groups taking the new and old forms in the fall of the year.

The results of the Cook, Eignor, and Schmitt (1988) study indicated that (1) matching on a set of common items provided greater agreement among the results of the various equating procedures studied, (2) for all equating procedures, the results of the common-items matched group equating agreed more closely with the criterion equating than did the results of the unmatched fall and spring equatings, and (3) matching on the selected questions from the SDQ did not improve results over the fall-spring random group equatings, possibly because a large number of examinees did not respond to the questions. The results of matching on scores on the common items provided an optimistic view about the possibility of being able to introduce new forms of the Biology Test at administrations where comparable old-form data did not exist.

The purposes of the present study were twofold: (1) to determine whether the improved fall-spring equating results brought about in the Cook, Eignor, and Schmitt study by matching on scores on common items generalize to ATP Achievement Tests in other content areas besides Biology, and (2) to continue to investigate the possibility of using covariates other than scores on an internal set of common items for matching purposes. Achievement Tests in Mathematics Level II, American History and Social Studies, Chemistry, and French were analyzed in the present study. Results from the previous study of Biology are included in this paper for comparative purposes. However, in the present study, IRT curvilinear true-score equating has been replaced by another curvilinear equating technique based on observed scores, frequency estimation equipercentile equating (see Angoff 1984, p. 113). Finally, for the French Test only, matched sample equatings were also performed using responses to selected questions on the SDQ (the old, not the revised version), to see whether the inadequate results noted for the Biology Test were replicated with these data. In addition, for the French Test only, responses to a question on amount of coursework contained on a Background Questionnaire (BQ) that appears with that test were used to match samples on ability. The French Test was the only test in the set being studied for which suitable (for matching) BQ and SDQ response data had been collected.

METHODOLOGY

Tests, Test Forms, and Samples

All tests in this study are formula scored with a correction for random guessing. The Biology, Mathematics Level II, American History and Social Studies, and Chemistry Tests all contain five-option multiple-choice questions, whereas the French Test contains four-option multiple-choice questions. Table 1 contains the numbers of items in the new and

old forms of each of the five tests in the study, along with the numbers of items in the common-item sets. It also contains information on when the new and old forms were administered. Formula scores on each of the Achievement Tests under study are transformed to scaled scores on a 200–800 scale, typically using conventional linear and equipercentile equating methods. A separate 200–800 scale exists for each test, and this scale is used for score reporting purposes.

Following is a description of the samples used in the various equatings for each test. An outline of the content and skills tested by each of the five Achievement Tests used in this study is found in Table 2. Each of the test forms for each of the tests was developed in accordance with these content specifications.

Biology

Random samples of approximately 2,500 examinees were selected from the total groups taking the new form at the fall administration and the old form at the spring and fall administrations. In addition, two matched samples from the spring old-form group were selected. One of these samples was selected in a nonrandom fashion to match the distribution of scores on the common items of the fall new-form sample. The other old-form matched sample was selected to match fall new-form group-cell frequencies in a bivariate cross-tabulation of responses to two questions from the SDQ, one question having to do with self-reported grades in biological sciences and the other with self-reported scientific ability (in comparison with others in the same age group). In forming both old-form matched samples, students were selected from the total spring old-form group ($N = 23,369$).

Figure 1 presents frequency distributions for the 36 common items administered to the fall new-form sample and the spring old-form total group. Figure 1 also contains the frequency distribution for the spring old-form matched group, where the matching variable was the score on common items. As can be observed in the column of percentages below, the spring old-form total group had a higher proportion of scores at the top of the 36-item range (approximately 50 percent scoring below 23). The fall new-form scores are distributed across a wider range of the common-item scores with approximately 50 percent scoring below 18. Because of the size of the spring old-form total group ($N = 23,369$), it was possible to create a spring old-form matched group that matched exactly the frequency distribution of the fall new-form group.

Figures 2 and 3 present SDQ question 15 (self-reported grades in biological sciences) and question 58 (self-reported scientific ability), respectively. The manner in which responses were recoded for cross-tabulation purposes and the correlations between responses on these questions and Biology Achievement Test scores for the fall new-form and spring old-form groups are reported in each of these figures. The correlations for each of these two questions are in the low to mid-40s, with slightly higher correlations for the spring old-form group.

Cross-tabulations of the recoded responses to SDQ questions 15 and 58 are displayed in Figure 4. The frequency in each cell of the cross-tabulation for the spring old-form group was matched to the corresponding cell frequency in the fall new-form group. Again, because of the size of the spring old-form total group, it was possible to create a spring old-form matched group that matched exactly, cell by cell, the bivariate cross-tabulation for the fall new-form group. Approximately half of the fall new-form group and of the spring old-form group (44 percent and 49 percent, respectively) failed to respond to either one of the two questions.

Formula-score summary statistics for the total tests and the 36-item common-item set for the five individual samples are presented in Table 3. The mean of the common-item set is represented as a percentage of the maximum possible score to provide an indication of the ability level of the five samples that is not dependent on the number of common items. The fall new- and old-form random groups and the matched common-items spring old-form group are closely matched in ability. The spring old-form random group had a considerably higher mean score on the common-item set, and the mean for the matched SDQ spring old-form group varied only slightly from the mean for this group.

Mathematics Level II

Random samples of approximately 2,000–2,300 examinees were selected from the total groups taking the new form at the fall administration and the old form at the spring and fall administrations. In addition, a matched sample from the spring old-form group was created. This sample was selected in a nonrandom fashion to match the distribution of scores on the common items of the fall new-form sample. In forming this old-form matched sample, students were selected from the total spring old-form group available in the data base ($N = 2,009$). Because of the size of the spring old-form total group, a spring old-form matched group could not be created to match exactly the frequency distribution of scores on the common items of the fall new-form group. The matched old-form group was created by proportional sampling in a manner such that the percentage below at each common-item score for the spring old-form matched group matched as closely as possible the corresponding percentage below at each score point for the fall new-form group.

Figure 5 presents frequency distributions for the 19 common items administered to the fall new-form group and the spring old-form total group. As can be observed in the column of percentages below, a higher proportion of the spring old-form total group scored at the top of the 19-item range (approximately 50 percent scoring at or above 10).

Formula-score summary statistics for the two total tests and the 19-item common-item set for the four individ-

ual samples are presented in Table 4. The fall new-form and old-form random groups and the matched common-items spring old-form group were closely matched in ability, as measured by the common items, while the spring old-form random group obtained a higher score on the common-item set.

American History and Social Studies

Random samples of approximately 2,000–2,100 examinees were selected from the total groups taking the new form at the fall administration and the old form at the spring and fall administrations. In addition, a matched sample from the spring old-form group was created. This sample was selected in a nonrandom fashion to match the distribution of scores on the common items of the fall new-form sample. In forming this old-form matched sample, students were selected from the total spring old-form group available in the data base ($N = 2,031$). Like Mathematics Level II, the size of this spring old-form total group precluded an exact matching of frequency distributions on the common items, and hence, proportional sampling (matching relative frequencies as closely as possible) was used to create the spring old-form matched group.

Figure 6 presents frequency distributions for the 20 common items administered to the fall new-form sample and the spring old-form total group. As can be observed in the column of percentages below, a higher proportion of the spring old-form group scored at the top of the 20-item range (approximately 50 percent scoring at or above 11).

Formula-score summary statistics for the total tests and the 20-item common-item set for the four individual samples are presented in Table 5. The fall new-form and old-form random groups and the matched common-items spring old-form group were closely matched in ability, as measured by the common items, while the spring old-form random group obtained a higher score on the common-item set.

Chemistry

Random samples of approximately 2,000–2,200 examinees were selected from the total groups taking the new form at the fall administration and the old form at the spring and fall administrations. As with the other tests, a matched sample from the spring old-form group was again created. This sample was selected in a nonrandom fashion to match the distribution of scores on the common items of the fall new-form sample. In forming this old-form matched sample, students were selected from the total spring old-form group available in the data base ($N = 2,206$). As was the case for the last two tests discussed, proportional sampling was again used in an attempt to match relative frequency distributions on the common items, resulting in the spring old-form matched group.

Figure 7 presents frequency distributions for the 20 common items administered to the fall new-form sample and the spring old-form total group. As can be observed in the column of percentages below, a higher proportion of the spring old-form group scored at the top of the 20-item range (approximately 55 percent scoring at or above 11).

Formula-score summary statistics for the two tests and the 20-item common-item set for the four individual samples are presented in Table 6. The fall new-form and old-form random groups and the matched common-item spring old-form group were closely matched, whereas the spring old-form random group obtained a higher score on the common-item set.

French

Random samples of approximately 6,100 examinees were selected from the total groups taking the new form at the fall administration and the old form at the spring and fall administrations. In addition, three matched samples from the spring old-form group were selected. One of these samples was selected, as was the case for the tests previously discussed, in a nonrandom fashion to match the distribution of scores on the common items of the fall new-form sample group. The second old-form matched sample was selected to match fall new-form group-cell frequencies in a bivariate cross-tabulation of responses to two questions from the SDQ, one having to do with years of study in foreign languages and the other with self-reported grades in foreign languages. The third old-form sample was selected to match fall new-form group frequencies of response to nine categories of a question on the BQ, concerning the manner in which candidates obtained their knowledge of French. In forming these old-form matched samples, students were selected from the total spring old-form group available in the data base ($N = 6,086$). As with previous examinations, proportional sampling (matching of relative frequencies) was used to create each of the spring old-form matched groups.

Figure 8 presents frequency distributions for the 21 common items administered to the fall new-form sample and the spring old-form total group. As can be observed in the column of percentages below, a higher proportion of the spring old-form group scored at the top of the 21-item range (over 50 percent at or above 9).

Figures 9 and 10 present SDQ question 8 (years of study of foreign languages) and question 14 (self-reported grades in foreign languages), respectively. The manner in which responses were recoded for cross-tabulation purposes and the correlations between responses on these questions and French Achievement Test scores for the fall new-form and spring old-form groups are reported in each of these figures. The correlations of Achievement Test scores with each of these two questions are in the range from .27 to .38, with the spring old-form group obtaining the higher correlations (.32 and .38).

Cross-tabulations of the recoded responses to SDQ questions 8 and 14 are displayed in Figure 11. The frequency in each cell of the cross-tabulation for the spring old-form total group was matched to the corresponding cell frequency in the fall new-form group. Exact matching of

cell frequencies was possible for all but one cell of the bivariate cross-tabulations. Approximately 52 percent of the fall new-form group and 30 percent of the spring old-form group failed to respond to either one of the questions.

Figure 12 presents the background having to do with how candidates obtained their knowledge of French. Figure 13 presents the distributions of responses to the BQ questions used in matching the spring old-form group to the fall new-form group. Proportional sampling was again used to form the spring old-form matched group.

Formula-score summary statistics for the two tests and the 21-item common-item set for the six individual samples are presented in Table 7. The fall new-form and old-form random groups and the matched common-items spring old-form group were closely matched, whereas the spring old-form group and the matched SDQ and matched BQ spring old-form group's scores on the common items were higher. The matched SDQ group varied only slightly from the spring old-form random group, whereas the matched BQ group was the most able of all samples.

Conventional Item Statistics (Equated Deltas)

The measure of item difficulty used in this study is the delta index, which is a transformation of the percentage or proportion of the group who answered the item correctly (see Hecht and Swineford 1981 or Henrysson 1981). Since deltas are direct transformations of proportion correct, the observed statistics (observed deltas) are dependent on the ability level of the group responding to the items. The solution to this dependency problem is to use item difficulties that are equated to some common scale (equated deltas).

Equating item difficulties requires establishing a base scale on an appropriate group. The relationship between observed deltas and deltas on the base scale can be expressed as an equating transformation of the form $\triangle_e = A(\triangle_o) + B$, where $\triangle_e$ is the equated delta value on the base score, $\triangle_o$ is the observed delta value to be transformed, and A and B are parameters estimated from the observed (for new-form sample) and equated (for old-form sample) deltas for the common items. The resulting transformation is then applied to all items in the test. Once the base scale is established, and once all observed deltas are equated to it, equated deltas can then be used to compare item difficulties that are not dependent on the different ability levels of the varying groups taking the items.

Of particular interest in this study are the relationships, for each test, between the equated deltas for the common items in the fall new-form random group and each of the various old-form groups formed for each of the tests.

Score Equating Methods

The five score-equating methods that were used in this study are (1) Tucker, (2) Levine equally reliable linear, (3) Levine unequally reliable linear, (4) chained equipercentile, and (5) frequency estimation equipercentile curvilinear methods. Each of these methods defines equated scores in a slightly different fashion.

Linear equating procedures used in this study employed the Tucker and the Levine equally and unequally reliable models (Angoff 1984, pp. 109–115). For these models, scores on the common items are used to estimate performance of the combined group of examinees on both the old and new forms of the test, thus simulating by statistical methods the situation in which the same group of examinees takes both forms of the test.

Chained equipercentile curvilinear equating establishes equivalency for scores on each total test and associated common-item set first by equipercentile methods separately within each group (Design V in Angoff 1984, p. 116). Scores on the two test forms are then said to be equivalent if they correspond to the same score on the common-item set.

Smoothing of equipercentile results can take place at two different points. Practitioners can either smooth the frequency distributions to be used in the equating or smooth the results obtained from the equating (see Fairbank 1987). In this study, a univariate smoothing technique, called "Tukey-Cureton seven-point smoothing" (Cureton and Tukey 1951; see also, Angoff 1984, p. 12) was applied to the frequency distributions to be used in the equipercentile equatings carried out through an anchor test.

Frequency estimation equipercentile curvilinear equating follows the logic of a single-group equating in that it equates new and old forms of a test on the basis of the total test-score distributions of the two forms in a single group of examinees. Usually, this group is the combined group formed from the groups taking the two tests to be equated. The total test-score distributions are estimated distributions, however, rather than observed distributions. These distributions are estimated using information contained in the cross-tabulations of common-item scores and total test scores. (See Angoff 1984, p. 113, for a further discussion of this procedure.)

For the frequency estimation equatings in this study, smoothing techniques were used at two points. A bivariate smoothing technique, attributable to Rosenbaum and Thayer (1987) and referred to as Model 4 in Rosenbaum and Thayer (1986), was used to smooth the joint distributions of total test scores and anchor-test scores in the two separate groups. The "Tukey-Cureton seven-point smoothing" technique was later applied to the univariate estimated total-test frequency distributions prior to equating.

Criterion Measures

Criterion measures for item analyses and score equating results are described in this section. For each test being studied, item analyses results were evaluated by comparing correlation coefficients and plots of difficulty indices (equated deltas) for the common items for the different sampling

combinations (fall-fall random groups, fall-spring random groups, fall-spring matched groups common items, fall-spring matched groups SDQ [Biology and French only], and fall-spring matched groups BQ [French only]). The fall-fall random groups' correlation coefficient and delta plot results were used as criteria to evaluate the results for the other sample combinations. The fall-fall random sample combination was chosen as the sample used to define the criteria because it provides equating results obtained under what has traditionally been considered to be the most satisfactory operational conditions.

Based on the results of the delta equatings for each new-form and old-form sample combination for each test, the new-form total-test equated delta mean (an indicator of total-test difficulty) and standard deviation were calculated. If the individual deltas are stable across new-form and old-form sample combinations, then the estimated total-test equated delta means and standard deviations should be similar. The fall-fall random groups estimated total-test equated delta mean and standard deviation were used for each test as the criteria in the comparisons.

For each test, score equating results for the linear equatings—Tucker, Levine equally reliable, Levine unequally reliable—and the nonlinear equatings—chained equipercentile and frequency estimation equipercentile—were compared by evaluating the within-method differences between the fall-spring random or matched groups equatings and the fall-fall random criterion equatings. In addition, for each test, across-method comparisons were made by examining the differences of each method and sampling combination from the fall-fall random groups Tucker criterion equating (the equating method used to report scores). In order to judge the importance of the results for the various models and sample combinations, standard errors were computed for the fall-fall random Tucker criterion equating. Confidence bands of ± 2 standard errors were plotted around the Tucker criterion on difference or residual plots that were used to evaluate selected equating results. The computer program AUTEST (Lord 1975) was used to compute the standard errors.

RESULTS AND DISCUSSION

Item Analyses

The results of the conventional item analyses for the five tests analyzed in this study are presented in Tables 8 through 12 and Figures 14 through 18.

Biology

An examination of the correlation coefficients in Table 8 obtained for the item difficulty indices (equated deltas) computed using the different sample combinations taking the Biology Achievement Test indicates that these coefficients range from .73 to .99. The highest correlation coefficient obtained (.99) was for delta values computed using the samples providing the criterion equating results (i.e., the random samples drawn from the fall new-form and fall old-form populations. The lowest correlation coefficient (.73) was obtained for deltas computed using the random samples selected from the fall new-form/spring old-form populations prior to any attempt to match the two groups on ability level. The correlation coefficients observed for deltas computed using samples obtained by the two matching procedures were quite similar (.79 and .77 for the groups matched using common items and responses from the SDQ, respectively).

Remember that the two matching procedures were not equally successful. As mentioned in the methodology section, matching on scores obtained on the common items produced an old-form matched sample that was close in ability to the new-form sample. On the other hand, matching the two groups using responses to the selected SDQ questions did not provide samples that were very similar in ability level. As a matter of fact, matching using SDQ responses appears to have accentuated the disparity between the ability levels of the two groups. Given that one matching procedure was so much more effective than the other, it was expected that the relationship between item-difficulty indices obtained for the better matched groups would be demonstrated by a correlation coefficient that was a good deal higher than that obtained for the deltas computed using the poorly matched samples (matched using SDQ responses) and the unmatched fall new-form and spring old-form groups.

Further insight into the behavior of the item difficulty indices can be obtained by examining plots of the equated delta values for the 36 common items administered to the four sample pairings (fall-fall random groups, fall-spring random groups, and the fall-spring groups matched on common items and SDQ responses). It can be seen, from examination of the plots presented in Figure 14, that for all sample combinations (with the exception of the fall-fall random groups combination, which provided the criterion score equating results), considerable scatter can be observed on either side of the solid diagonal line. The correlation coefficients shown under the plots, and in Table 8, are reflective of this degree of scatter.

Again, note that neither of the matching techniques reduced substantially the scatter near the diagonal line for the equated delta values computed using the new- and old-form samples. As mentioned earlier, matching samples on ability level using the two SDQ questions was not successful, most likely because of the large number of examinees choosing not to respond to the questions. It was expected, however, that once examinees were matched on ability level on the common-item set, delta values computed using the matched groups would show a close relationship, similar to that exhibited by the fall new-form and fall old-form random samples used to provide the criterion equatings. The fact that scatter remains in this delta plot, despite the matching

of the new- and old-form samples, is troublesome. A likely explanation is that the data are multidimensional, with the common-item portion of the tests measuring different abilities for the spring and fall groups, and matching on common-item scores does not remove this multidimensionality from the data. In addition, matching on a fallible measure, such as scores on a set of common items that contain some degree of measurement error, may not provide a match on true ability.

Table 8 also contains estimated total-test equated delta means and standard deviations for the new form of the test, based on the delta equatings obtained using the common items administered to the various samples. It is interesting to note that neither the means or the standard deviations of equated deltas appear to be affected by the groups used to perform the delta equatings. The equated delta means can be interpreted as indicating that the new form is estimated to be of similar difficulty level for all sample combinations, despite the differences in ability level that exist for two of the combinations.

Mathematics Level II

Table 9 and Figure 15 contain information summarizing the results of the item analyses of the Mathematics Level II Test. Correlations among the 19 common items appearing in the new and old form of the test are quite high for all sample combinations (.98–.99), and the correlation does not appear to be substantially affected by the differences in ability levels of the spring old-form and fall new-form groups.

The plots of equated delta values provided in Figure 15 reflect the high correlation coefficients for the common items given to the various sampling combinations. It can be seen, from examination of the plots shown in Figure 15, that the item-difficulty indices exhibit very little scatter about the diagonal line drawn on these plots.

It would appear that, although examinees differ in level of ability for the fall new-form and spring old-form random groups sample combination, the rank order of the difficulty indices for the common items remains similar across all sample combinations. The total-test equated delta means and standard deviations (shown in Table 9) remain consistent across the various sampling combinations.

American History and Social Studies

The information presented in Table 10 and Figure 16 summarizes the item analysis results for the American History and Social Studies Achievement Test. Correlation coefficients for the 20 common items given to the fall new-form, fall old-form, and fall new-form/spring old-form random and matched groups vary between .93 and .94.

Plots of equated delta values shown in Figure 16 are reflective of the correlation coefficients provided in Table 10, showing some degree of scatter about the diagonal line. It would appear that, to some extent, the American History and Social Studies Test provides common-item data that are somewhat intermediate in terms of scatter about the diagonal line to that collected for the Biology Test and the Mathematics Level II Test. The common-item portion of the tests may possibly be measuring slightly different constructs for the new- and old-form groups (as exhibited by the scatter of the common items). This observation seems to hold equally for the fall-fall random groups combination as well as for the two fall-spring combinations.

It should be noted that the total-test equated delta means and standard deviations shown in Table 10 are quite consistent across the various sampling combinations.

Chemistry

Item analysis results for the Chemistry Achievement Test are shown in Table 11 and Figure 17. Examination of the correlation coefficients for the 20 common items appearing in the new and old Chemistry Test forms indicates some effect of the differences in the ability level of the fall and spring groups on the relationship between the common items contained in the two forms. However, total-test equated delta means and standard deviations remain quite consistent across the different sampling combinations.

Plots of the equated delta values for the Chemistry Test common items are shown in Figure 17. Examination of these plots indicates slightly more scatter of equated delta values about the diagonal line for the items taken by the fall-spring combination than for those taken by the fall-fall combination. Matching the new- and old-form groups on ability, on the basis of the common-item scores, has little or no effect on the scatter of the equated delta values.

French

The final set of item analysis results, those obtained for the French Achievement Test, are provided in Table 12 and Figure 18. The information provided in Table 12 indicates that the correlation coefficients between the 21 common items contained in the new and old forms of the test, as well as the total-test equated delta summary statistics, are remarkably similar across four of the sampling combinations. The fall-spring matched on BQ combination, however, provided a somewhat different set of total-test equated delta summary statistics.

The plots of equated delta values shown in Figure 18 illustrate graphically the close agreement of the equated delta values for the various sampling combinations, in spite of the considerable differences in ability levels of the fall and spring groups. The relationship between the item-difficulty values seems not to be particularly affected by differences in group ability.

Comparison of Item Analysis Results Across Tests

To summarize, item analysis results (correlations between equated delta values) appear to be most affected by the differences in ability exhibited in the fall-spring sampling combinations for the Biology Achievement Test and least affected for the French and Mathematics Level II Tests. Item analysis results for the American History and Social Studies

Test indicate only a slight decrease in equated delta correlations for the fall-spring random groups results compared with results obtained for the fall-fall random groups combination. The results of the item analyses carried out for the Chemistry Test indicate the next largest drop in correlation between equated delta values obtained for the fall-spring random groups as compared with the results obtained by the fall-fall random groups. The largest drop for any test studied, as mentioned previously, was observed for the Biology Test. These results make sense intuitively, in that science tests are more closely tied to the content of a single course than are tests in mathematics or a foreign language, which measure skills obtained from several courses on the topic. Therefore, one might expect, because of recency of instruction and, possibly, some degree of forgetting, that differences would occur in the behavior of common items contained in science-test forms when given to students just completing a course in the subject matter (a spring group) and when given to students who may not have studied the subject matter for some time (a fall group).

Equatings

Scaled-score summary statistics resulting from the experimental equatings performed for the five tests are summarized in Tables 13–17. Insight into the performance of the various equating methods when applied to the different sample combinations for the five tests can be obtained through a review of the equating plots provided for these tests in Figures 19–33.

Biology

Examination of the information presented in Table 13 indicates that the scaled-score summary statistics obtained for all of the Biology Test equatings performed using the random samples selected from the fall new-form and fall old-form combination are very similar. Remember that the criterion equating for the study was specified as being the equating actually used to report scores for the fall new-form group. Scores were reported for Biology Test takers using the fall-fall random group Tucker results. However, it is clear from examination of the information provided in Table 13 that all procedures used with this sampling combination would have provided similar total-group scaled-score means and standard deviations.

It is important to note the differences in equating results between those obtained from the fall-fall random groups and those obtained from the groups that were randomly selected from the fall new-form and spring old-form populations. It is clear from the summary statistics provided in Table 13 that the differences in ability levels of the new- and old-form groups taking the Biology Test (see Table 3) affect the equating results obtained by the four equating procedures differentially. An interesting point to note, when examining these results, is how little the Tucker and frequency estimation equipercentile procedures are affected by the differences in ability levels of the new- and old-form samples. Comparison of the scaled-score summary statistics obtained for these experimental conditions with the results obtained for the Tucker procedure applied to the fall-fall random groups sampling combination shows very little difference in these results. On the other hand, the other three equating procedures investigated (Levine equally reliable, Levine unequally reliable, and chained equipercentile) appear to be quite affected by the ability differences represented by the fall new-form and spring old-form random groups combination. The equating results that appear to be the most affected are those based on the Levine equally and unequally reliable methods. A point to note is the inconsistency of the results of the procedures for this particular new-form and old-form pairing. It would be difficult to identify the most appropriate set of results to use for score reporting in the absence of some criterion.

Examination of the information presented in Table 13 for the equatings based on the fall new-form and spring old-form groups matched using scores obtained on the common items indicates remarkable consistency among the summary statistics obtained using all five equating procedures. One important point to note is that, although matching provides a definite improvement (in relation to the criterion scaled-score mean and standard deviation) over the fall-spring random groups equating results for the two Levine procedures and chained equipercentile procedure, this improvement is not noted for the Tucker and frequency estimation equating procedures. The summary statistics produced by the Tucker and frequency estimation equatings based on the fall-spring random groups are closer to those produced by the criterion equating than are the summary statistics resulting from the Tucker and frequency estimation equatings based on the fall-spring matched (common items) groups.

As expected, the equating results based on the Biology fall new-form and spring old-form groups matched on the responses to the two SDQ questions produced summary statistics that were discrepant from the criterion equating summary statistics. Several points should be noted when examining these data. First, a good deal of inconsistency among the scaled-score summary statistics produced by the various equating procedures exists, with the results produced by the two Levine procedures and the chained equipercentile procedure providing lower scaled-score means than either the Tucker or frequency estimation equipercentile results. Using the Tucker results from the fall-fall random groups as a criterion, it is apparent that the two Levine procedures used with this sampling combination produce the results that are most discrepant from the criterion. The Tucker and frequency estimation equipercentile procedures produce results that most closely approximate the criterion results.

Plots presented in Figure 19 allow the comparison of the Biology Test equating results obtained by a particular equating procedure (e.g., Tucker) for the four Biology Test sampling combinations. Panel (a) contains results for the Tucker procedure; panel (b) contains results for the Levine

equally reliable procedure; panel (c) contains results for the Levine unequally reliable procedure; panel (d) contains results for the chained equipercentile procedure; and panel (e) contains results for the frequency estimation equipercentile procedure. Each panel contains plots of differences in scaled scores, with the fall-fall random groups equating results subtracted from the fall-spring random or matched groups equating results. In each case, the fall-fall results used for comparison are those obtained by the particular equating procedure that is being evaluated.

Examination of the results obtained in panel (a) of Figure 19 shows the close agreement among the Tucker results under all sampling conditions that was illustrated by the Tucker scaled-score summary statistics for the Biology Test found in Table 13. Only the residuals obtained from the comparison for the fall-spring groups matched on SDQ responses are greater than five scaled-score points, and this is found only for raw scores in the lower end of the distribution.

The results presented in panels (b) and (c) for the Levine equally and unequally reliable methods are similar to each other and quite different from those shown in panel (a). It is clear that for both Levine methods the only results that closely approximate the results obtained under the fall-fall random condition are the results obtained for the fall-spring groups matched on common items. For both Levine procedures, the fall-spring random groups and the fall-spring groups matched on SDQ responses produce equating results that are quite discrepant from the fall-fall random groups equating.

The results presented in panel (d) for the chained equipercentile procedure are difficult to interpret because of the irregularities in the raw to scale conversions. However, it can be seen that the results that most closely match those obtained for the fall-fall random groups combination are those for the fall-spring groups matched on common items.

The frequency estimation equipercentile equating results are summarized in panel (e) of Figure 19. An examination of these results indicates a degree of similarity to the results obtained by the Tucker linear procedure in that the closest results to the criterion equating (frequency estimation equipercentile for the fall-fall random groups in this case) are provided by the fall-spring random groups equating. The most discrepant results are produced by the fall-spring groups matched using information from the SDQ. Note that all results depart considerably from the criterion results for scaled scores corresponding to raw scores below 20; this is most likely partly due to instability in the equatings related to the small number of frequencies in this portion of the score distributions.

The Biology Test residual plots presented in Figure 20 have been developed to permit an examination of the agreement among the five equating procedures applied to a particular sampling combination. The plots shown in panel (a) of Figure 20 compare results for the Tucker, two Levine, chained equipercentile and frequency estimation equipercentile equating methods used with the fall-fall random groups samples. Panel (b) of Figure 20 summarizes the results of the fall-spring random groups equatings. The fall-spring groups (matched on common items) equatings are summarized in panel (c), and the equatings performed using the fall-spring groups (matched on SDQ responses) are summarized in panel (d). Note that the same criterion equating is used in each plot, that is, the results of the Biology Test Tucker fall-fall random groups equating. In addition, a confidence band of ± 2 standard errors has been plotted around the criterion equating in each plot. If the differences between the criterion and experimental results fall within this band, it can be concluded that these equating differences are no larger than what would be expected from sampling fluctuations (with respect to the criterion) alone.

Examination of the Biology Test equating-difference plots shown in panel (a) of Figure 20 indicates that all equating methods, applied to data obtained from the fall-fall random groups combination, produce results that are in fairly close agreement. Results of the two Levine procedures and frequency estimation equipercentile equating procedure agree very closely with those obtained using the Tucker model. The chained equipercentile results are generally close to the criterion results and do not show a tendency to consistently over- or underestimate equated scores.

Panel (b) contains the results of applying the five equating procedures to data obtained from the Biology Test fall new-form and spring old-form random groups sampling combination. Examination of the plots indicates that both the Tucker and frequency estimation equipercentile procedures provide results that agree in a reasonable fashion with the criterion results. The difference line for the Tucker equating falls within the confidence band plotted around the criterion equating results for the full range of raw scores. This is true also for the frequency estimation equipercentile results with the exception that discrepancies for this procedure fall out of the confidence band for low raw scores where there is little data. The Levine equally and unequally reliable and chained equipercentile procedures produce results that are quite discrepant from the criterion equating results. Both the Levine and the chained equipercentile procedures underestimate the criterion scores for most of the raw score distribution. The equipercentile results do, however, appear to more closely approximate the criterion results than do the results of the Levine procedures.

The results shown in panel (c) of Figure 20 summarize the equating differences obtained from a comparison of the results of the five equating methods used with the fall new-form and spring old-form groups (matched using scores on the common items) with results obtained for the Tucker model applied to the fall-fall random groups combination. It can be seen that most of the equating differences resulting from application of the five procedures fall within the confidence band plotted around the criterion equating results.

Panel (d) contains the residuals resulting from a comparison of the criterion equating results with the equating

results obtained from application of the five equating procedures to data from the fall new-form and spring old-form groups matched using responses to the SDQ questions. The pattern of residuals observed in panel (d) is similar to that previously examined for panel (b), with the exception that the equating differences are even more pronounced for the groups matched using SDQ responses. It is clear, from the information presented in panel (d), that the Tucker and frequency estimation methods provide results that agree somewhat with those obtained by the Tucker model applied to the fall-fall random groups combination. The equating procedures that provide unacceptably large residuals are the ones involving application of the Levine equally and unequally reliable and chained equipercentile models.

Figure 21 summarizes the results of the application of the five equating procedures to the four Biology Test sampling combinations. New-form scaled-score means are plotted for each equating-method and sampling combination. A number of points that have already been mentioned become even more apparent from an examination of the plots shown in Figure 21. First, it is clear that the methods that provide reasonably consistent results across the four new-form and old-form sampling combinations are the Tucker and frequency estimation equipercentile methods. The next most stable results are provided by the chained equipercentile procedure. The least stable results are provided by the Levine equally and unequally reliable procedures. These procedures appear to be severely affected by differences in group ability, as exhibited in the results obtained for the fall-spring random and fall-spring matched (SDQ) sampling combinations. Additional points worth noting are:

1. There is close agreement of the results of all the methods for the fall-fall random groups combination;
2. Matching on common items promotes agreement among the five methods, with all five producing results that slightly underestimate the criterion scaled-score mean; and
3. Better results (i.e, closer estimates of the criterion scaled-score mean) are obtained for the Tucker and frequency estimation equipercentile methods by using the fall-spring random groups combination and not matching.

For Levine equally and unequally reliable and the chained equipercentile procedures, matching on common items clearly provides better results.

The results of the application of the Tucker and frequency estimation equipercentile models to the various sampling combinations are interesting from several points of view. The most important point is that the scaled-score summary statistics produced by these methods applied to the fall new-form and spring old-form random groups combination very closely approximate the criterion summary statistics. This implies that, for this test, the Tucker and frequency estimation equipercentile methods are relatively unaffected by new-form and old-form sampling combinations that differ in ability level. This is certainly not true for the other three equating procedures applied to the Biology Test data. These procedures clearly appeared to be affected by the group ability differences displayed by the fall-spring random groups and fall-spring matched groups (SDQ) sampling combinations.

Mathematics Level II

The results of the experimental equatings carried out using data from the Mathematics Level II Test are summarized in Table 14 and Figures 22–24. Scaled-score summary statistics resulting from application of the five equating methods to this test appear in Table 14. Examination of the information presented in this table indicates that the scaled-score summary statistics obtained for all of the Mathematics Level II equatings performed using the random samples selected from the fall new-form and fall old-form combination are quite similar. Remember that the criterion equating for the study was specified as the fall-fall random groups Tucker equating actually used to report scores for the fall new-form group. However, it is clear that all procedures used with this sampling combination would have resulted in similar reported scores.

Differences between the results of applying the five equating procedures to the fall-fall sampling combination and to the fall-spring random groups combination are quite apparent from the scaled-score summary statistics presented in Table 14. The procedures that appear to be least affected by differences in group ability are the two Levine procedures, and the procedures most affected by group differences are the Tucker and frequency estimation equipercentile procedures.

The information presented in Table 14 for equatings based on fall-spring samples matched on their respective distributions of common-item scores are quite interesting. Matching provides similar results for all five equating procedures and results that provide an overestimate of the criterion scaled-score mean and standard deviation.

Further insight into the equating summary statistics presented in Table 14 for the Mathematics Level II Test can be obtained by examination of the equating-difference plots provided in Figures 22 and 23. Plots shown in Figure 22 provide comparisons of the results of a single procedure applied to the three sampling combinations. The plots shown in Figure 23 permit comparison of different equating procedures applied to a particular sampling combination.

Results contained in panel (a) of Figure 22 show the discrepancy among the Tucker results, with both the fall-spring random and fall-spring matched common-item results overestimating the criterion scaled scores produced by the Tucker fall-fall random groups equating. It is interesting to note that matching had little effect on the fall-spring results. The matched group equatings produced slightly low-

er scaled scores than the fall-spring random groups equating for the upper end of the distribution of raw scores and slightly higher scaled scores for the lower end of the distribution.

Information provided in panels (b) and (c) for the Levine equally and unequally reliable procedures, respectively, indicates that for both procedures the matched and random groups fall-spring combinations provide overestimates of the Levine equating results based on the fall-fall random groups combinations. For both procedures, matching, based on distributions of scores on the common-item set, provided results that overestimated scaled scores obtained for the fall-fall combination more than the results provided by the unmatched fall-spring combination.

Examination of the information presented in panel (d) indicates that the results obtained using chained equipercentile equating agree closely for both the fall-spring random and fall-spring matched common-item groups and that results of both of these procedures overestimate scaled scores obtained through application of the chained equipercentile procedure to data obtained for the fall-fall random groups combination. Finally, the results provided for the frequency estimation equipercentile procedure (panel (e) of Figure 22) are similar to those obtained for the other four procedures (i.e., both matched and unmatched results overestimate the equating results obtained from the fall-fall sampling combination).

The Mathematics Level II residual plots shown in Figure 23 permit comparisons of the various equating methods applied to a particular sampling combination with the criterion equating results (Tucker applied to the fall-fall sampling combination). The plots shown in panel (a) compare results for the Tucker, Levine equally and unequally reliable, chained equipercentile and frequency estimation equipercentile methods used with the fall-fall random groups samples. Close agreement among the various methods is exhibited for a raw-score range of approximately 5–40. The two equipercentile methods diverge from each other and from the linear criterion equating results for extreme raw scores. This is most likely due to the instabilities of both equipercentile procedures related to the small number of frequencies in the tails of the score distributions.

Examination of the information provided in panel (b) of Figure 23 reveals that the only procedures (applied to the fall-spring random groups combination) that provide scaled-score discrepancies within ± 2 standard errors of the criterion equating are the Levine equally and unequally reliable procedures. All other procedures appear to provide results that overestimate the criterion scaled scores through the middle portion of the score range. The scaled-score discrepancies plotted in panel (c) are quite similar to those shown in panel (b), indicating that matching using distributions of scores on the common items had little effect on the equating results. Exceptions are the results of the Levine equally and unequally reliable procedures. These procedures seem to have been made worse by the matching. For the Levine procedures, the matched results produce scores that are a considerable overestimate of the Mathematics Level II criterion scores.

Figure 24 summarizes the results of the application of the five equating procedures to the three Mathematics Level II sampling combinations. New-form scaled-score means are plotted for each equating-method and equating-sampling combination. From examination of Figure 24, it is apparent that all of the equating procedures are affected to some extent by the difference in group ability exhibited by the fall-spring random groups combination, with the Tucker and frequency estimation procedures affected a good deal more than the Levine procedures. It is also apparent that matching on the basis of common items, although it provides greater similarity among the scaled-score means obtained by the five methods, provides only slightly improved results for the Tucker and frequency estimation procedures and actually worsens the results obtained by the Levine and chained equipercentile procedures.

American History and Social Studies

Table 15 and Figures 25–27 contain the results of the equatings carried out for the American History and Social Studies Achievement Test. Table 15 contains scaled score summary statistics resulting from the application of the five equating methods to the three sampling combinations (fall-fall random, fall-spring random, and fall-spring matched using common items). Examination of the information provided in Table 15 indicates close agreement among the results of the five equating methods applied to the fall-fall random groups sampling combination. The criterion is the result of the Tucker fall-fall random groups equating used to report scores for the fall new-form group. However, as was the case for the Biology and Mathematics Level II Tests, it is clear that all equating procedures used for the fall-fall sampling combination would have resulted in similar scaled-score summary statistics.

The results of applying the five equating methods to the fall-spring random groups sampling combination demonstrate that the equating procedures most clearly affected by the differences in group ability are the Levine procedures. The Tucker and frequency estimation procedures behave similarly for the fall-spring random groups combination, each producing scaled scores that overestimate the criterion scaled scores. The equating procedure used with this test that appears to be the least affected by differences in group ability is the chained equipercentile procedure.

The scaled-score summary statistics resulting from the samples constructed by matching (based on the distributions of scores on the common items) presented in Table 15 show that matching provides greater agreement among the five equating methods applied to the American History and Social Studies Test. It should be noted that, in all cases, matching provides scaled-score summary statistics that

appear to overestimate the criterion scaled-score summary statistics.

Figures 25 and 26 provide equating-difference plots for the American History and Social Studies Test. Examination of the results contained in panel (a) of Figure 25 indicates the close agreement between the Tucker results obtained from the fall-spring random and fall-spring matched on common-items groups. It is clear that the Tucker results, applied to both sample combinations, overestimate criterion scaled scores in the upper end of the raw-score distribution and underestimate criterion scaled scores for the lower end of the distribution.

The information presented in panels (b) and (c) is indicative of the serious effect the differences in ability levels of the fall and spring groups have on the Levine equally and unequally reliable equating procedures applied to this test. (See Table 5 for a comparison of these differences.) Both the Levine procedures used with the fall-spring random groups sample combination provide an underestimate of the scaled scores obtained using data from the fall-fall random groups combination. These equating procedures applied to the fall-spring matched through common-items groups provide an overestimate of scaled scores in the upper end of the raw-score distribution and an underestimate of scaled scores in the lower end of the distribution when compared with the fall-fall random groups results.

Information provided in panel (d) of Figure 25 indicates the effect on the results of the chained equipercentile procedure of matching on the basis of the distribution of scores on the common items. It is quite clear that matching had a deleterious effect on this equipercentile procedure used with this particular test. Panel (e) contains the results of the frequency estimation equipercentile procedure. The plots shown in this panel indicate that the frequency estimation results are affected by the differences in group ability displayed by the fall-spring random groups and that matching on the basis of common-item score distributions scoring seems to slightly exacerbate this effect.

Figure 26 contains plots of equating results for the American History and Social Studies Test that allow comparisons of the five equating methods applied to a particular sampling combination. The results of the equatings shown in panel (a) of Figure 26 indicate that both curvilinear procedures, particularly the chained equipercentile procedure, provide results that fall slightly outside the ± 2 standard error bands for scores in the middle part of the distribution and again for scores in the extremes of the distribution.

Panel (b) contains the results of the five equating methods applied to the fall-spring random groups sampling combination. Examination of the plots contained in panel (b) shows the serious effect that differences in group ability, displayed by the fall-spring random groups combination, have on the Tucker and the two Levine equating procedures. Scaled scores resulting from the two curvilinear procedures fall mostly within the standard error bands provided for the criterion-equating results for raw scores ranging from about 10 to 60. Frequency estimation equating results show a tendency, as do results of the other methods, to seriously overestimate scaled scores obtained for raw scores in the upper end of the distribution.

Information provided in panel (c) of Figure 26 summarizes the results of the five equating procedures applied to the fall-spring groups matched using distributions of scores on the common items. The plots shown in panel (c) indicate that matching, using the frequency distribution of scores on the common-item set, brings the results of all four equating procedures closer together; in particular, the linear procedures now show fairly close agreement. It appears as though all the procedures provide scores that are a considerable overestimate of the criterion scores for the upper end of the raw-score distribution.

Figure 27 contains plots of American History and Social Studies Test new-form scaled-score means obtained for each equating-method and equating-sampling combination. It is clear from an examination of the scaled-score means that all equating methods, with the exception of the chained equipercentile method, are affected by the differences in group ability presented by the fall-spring random group sampling combination. The differences in sample ability level affect the Tucker and frequency estimation procedures in the opposite direction from the effect the differences have on the Levine equally and unequally reliable procedures. An additional point that is apparent from the plots is that matching, using scores on the common-item set, has a deleterious effect on all methods except the Levine equally and unequally reliable methods in that the resulting means are more discrepant from the criterion mean than the means resulting from no matching. Matching does provide results among the methods that are more in agreement.

Chemistry

The experimental equatings performed for the Chemistry Achievement Test are summarized in Table 16 and Figures 28 through 30. Table 16 contains scaled-score summary statistics resulting from application of the five equating methods to the three new-form and old-form sampling combinations (fall-fall random, fall-spring random, and fall-spring matched on common items).

The information shown in Table 16 indicates that the five equating methods provided results in close agreement for the fall-fall random groups sampling combination. Results for the equatings using the fall-spring random groups combination differed considerably from the criterion scaled-score means (the fall-fall Tucker results used to report scores for the fall new-form group). As was the case for the results of the equatings for the tests previously discussed, Tucker and frequency estimation equipercentile equating methods provide similar results for this sampling combination, and the Levine equally and unequally reliable and chained equipercentile procedures provide results that differ somewhat from each other and from the other two equating procedures. Examination of the results obtained for the five

equating methods applied to the Chemistry Test fall-spring groups matched using the common-item set shows close agreement among all five equating procedures; however, the discrepancies between the matched-group means and the criterion mean are quite pronounced, more so than for any test discussed so far.

Figure 28 contains plots of equating differences for the Chemistry Test that parallel those described for the three tests that have been previously discussed. The plots presented in Figure 28 have been developed to permit comparisons of the manner in which a particular equating method performs across the three sampling combinations.

Examination of the information presented in panel (a) of Figure 28 shows that the Tucker method applied to the Chemistry Test data is affected by the differences in group ability exhibited by the fall-spring random groups combination. Furthermore, matching the new- and old-form groups using distributions of scores on the common-item test has little effect on the Tucker results. Tucker results for both fall-spring combinations were quite similar.

The information provided in panels (b) and (c) indicates that the Levine equally and unequally reliable results are also affected by differences in group ability and provide a considerable overestimate of scaled scores (when compared with the results of the fall-fall sampling combination) in the upper end of the score distribution. Matching on the basis of distributions of scores on the common items provides peculiar results in that the matched-group equating provides scores similar to those obtained from the unmatched fall-spring group for the upper end of the score distribution and scores that are more of an overestimate for the lower end of the score distribution. In general, matching seems to have a deleterious effect on both Levine procedures.

Information provided in panels (d) and (e) (for the chained equipercentile and the frequency estimation equipercentile procedures, respectively) is quite similar. In general, both procedures appear to be affected by the differences in group ability displayed by the fall-spring random groups combination. And, as was the case for the Tucker and Levine procedures used with this test, matching on the basis of common items provides somewhat worse results for both procedures but does not have as serious an effect on the results as that observed for the Levine procedures.

Figure 29 contains equating-difference plots that have been developed to provide comparisons of the different equating procedures applied to a particular sampling combination.

Examination of the information provided in panel (a) of Figure 29 shows that, for the fall-fall random groups combination, the equating results provided by the two Levine procedures do not differ substantially from the Tucker criterion results. The two curvilinear procedures, chained equipercentile and frequency estimation equipercentile, also appear to agree fairly well with the criterion results through the midportion of the score range. The plots shown in panel (b) of Figure 29 illustrate dramatically the effect that the differences in group ability have on the five equating methods used with the Chemistry Test. For the fall-spring random groups combination in panel (b), none of the procedures provide results that are within two standard errors of the criterion scaled scores. However, the effect of matching on the common-item scores also supplies extreme results, as illustrated by the plots shown in panel (c), with all methods providing similar overestimates of the criterion scaled scores.

Further insight into the Chemistry Test equating results can be obtained by an examination of the scaled-score means for the various equating procedures and sampling combinations that are provided in Figure 30.

Examination of the information provided in Figure 30 shows the serious effect on all five equating procedures of the ability differences exhibited by the fall-spring random groups combination. In this case, the Tucker and frequency estimation procedures appear to be the most affected, whereas the Levine equally and unequally reliable procedures are the least affected. Matching on the basis of the common-item set provides scaled-score means for the two Levine and chained equipercentile procedures that are considerably more discrepant from the criterion scaled-score mean than those provided by the unmatched fall-spring random groups equating. Matching also appears to provide slightly more discrepant means (as compared with the unmatched fall-spring means) for the Tucker and frequency estimation equipercentile procedures.

French

Scaled-score summary statistics resulting from the experimental equatings for the French Test are provided in Table 17. Examination of the information provided in Table 17 indicates that all the French Test equatings performed using the random samples selected from the fall new-form and old-form combination are quite similar. Again, the criterion equating for the French Test is the equating actually used to report scores for the fall new-form groups (i.e., the fall-fall random groups Tucker results). It is important to note the differences obtained by applying the five equating procedures to the groups randomly selected from the fall new-form and spring old-form populations. It is clear, from the summary statistics provided in Table 17, that the differences in ability levels of the new- and old-form groups taking the French Test have a substantial effect on the equating results for all procedures, with the Levine procedures affected the least.

Examination of the information presented in Table 17 for the equatings based on the French Test fall new-form and spring old-form groups matched using scores obtained on the common items indicates, as was the case for all tests discussed thus far, remarkable consistency among the scaled-score summary statistics obtained by application of the five equating procedures. An important point to note is that, although matching promotes consistency among the

results of the five equating methods, all five sets of summary statistics provide overestimates of the criterion scaled-score mean and standard deviation.

As expected, the equating results based on the French Test fall new-form and spring old-form groups, matched using responses to the two SDQ questions or on the single BQ question, produced scaled-score summary statistics that were discrepant from the criterion equating summary statistics. Matching on the basis of SDQ responses appears to have provided summary statistics, particularly means, that are very similar to those obtained for the unmatched fall-spring random groups combination. Matching on the basis of BQ responses appears to have exacerbated the situation by providing statistics even more discrepant from the criterion equating results than those obtained using the fall-spring random groups equatings.

Further insight into the equating results presented in Table 17 can be obtained by examining the equating-difference plots shown in Figure 31. Similar to the other tests discussed thus far, the plots shown in Figure 31 allow comparison of the results of the application of the particular equating procedure across the five sampling combinations. Remember that, in each case, the fall-fall results used for comparison are those obtained by the particular equating procedure that is being evaluated.

Examination of the results presented in panel (a) of Figure 31 shows the close agreement among the Tucker results for all the fall-spring sampling combinations (with the exception of the results based on the BQ matching) and how these results overestimate the criterion scaled scores through most of the raw-score range, but particularly in the upper end of the raw-score distribution. The results presented in panels (b) and (c) are somewhat similar to the Tucker results presented in panel (a). In general, the fall-spring random and matched Levine results (equally and unequally reliable) have a tendency to overestimate scores (when compared with the results of the fall-fall Levine equatings), particularly in the upper end of the score distribution. It is important to note how matching on the basis of scores obtained on the common-item set appears to worsen the situation.

The results presented in panel (d) for the chained equipercentile procedure indicate that all equatings based on the fall-spring sampling combinations (both matched and unmatched) have a tendency to overestimate scaled scores when compared with the fall-fall random groups results. This trend is also apparent for the frequency estimation equipercentile results summarized in panel (e). In both cases, the equating results based on matching using scores on the common items appear to provide scaled scores that are more of an overestimate of the fall-fall results than those obtained using the unmatched fall-spring random groups samples.

The French Test residual plots shown in Figure 32 have been developed to permit the examination of the agreement among the five equating procedures applied to a particular sampling combination. The plots shown in panel (a) compare results for the five equating procedures applied to the fall-fall random groups samples. Those shown in panel (b) are for the equating procedures applied to the fall-spring random groups combination. Panel (c) contains plots pertinent to the evaluation of the results of equatings based on groups matched using common-item scores; panel (d) contains equating results for groups matched on the basis of the two SDQ questions and panel (e), equating results for groups matched on the basis of their responses to the single BQ question.

Examination of the difference plots shown in panel (a) of Figure 32 indicates fairly close agreement of the results obtained for the five procedures with the criterion scaled scores; however, it should be noted that both equipercentile procedures have a tendency to underestimate criterion scaled scores in the middle of the distribution and overestimate these scores in the ends of the raw-score distribution. Panel (b) of Figure 32 contains equating discrepancies for the five methods applied to the fall-spring random groups combination. It is clear, from examination of the plots shown in panel (b), that all procedures had a tendency to overestimate the criterion scaled scores in the upper end of the raw-score distribution.

Panels (c), (d), and (e) contain the results of matching the fall new-form and spring old-form groups in ability level on the basis of scores on a common item set, responses to the SDQ questions, or responses to the BQ questions. It is clear, from an examination of the results provided in these three panels, that matching, using any of the three covariates, did not provide satisfactory results. Matching on the basis of common-item scores, as seen in panel (c), provides consistent results across equating methods, but these results also provide consistent overestimates of the criterion scaled scores.

Figure 33 summarizes scaled-score means resulting from the application of the five equating procedures to the five sampling combinations used for the French Test. Examination of the information presented in Figure 33 shows several trends that have been observed from the other Achievement Tests that have been discussed so far. First, equatings based on groups that are similar in ability level provided consistent results across the five methods examined. Second, equatings based on groups differing in ability level provide different results depending on the equating procedures that are used. In the case of the French Test, the Tucker and frequency estimation procedures appear to be the most affected by differences in ability levels, whereas the Levine procedures appear to be the least affected.

Comparison of Equating Results Across Tests

It is important to compare the results provided by the five equating methods applied to the various sampling combinations for the five tests used in this study. For this comparison, the results of matching using responses to the SDQ or BQ are not discussed. Basically, it was not possible to match groups on ability level using these covariates; therefore, including the equating results obtained for these

samples in a comparison of the results of equatings using matched versus unmatched samples would not contribute to the discussion.

Figures 21, 24, 27, 30, and 33 provide the most useful information for comparing equating procedures and sampling combinations across the five tests. As mentioned previously, several trends are worth noting when comparing scaled-score means contained in these figures. First, the most consistent result across all five tests is that when new- and old-form samples are similar in level of ability, either because they are randomly selected from similar populations or because they have been deliberately constructed to be the same (i.e., matched on a set of common items), all the equating procedures evaluated in this study provide similar results. Second, in almost all instances, matching the old-form group to the new-form group using an internal set of common items produces scaled-score means that differ from those obtained when random samples from fall populations are used.

A third observation is that differences in ability levels of new- and old-form samples affect equating procedures differently across tests. In the case of the Biology Test, the procedures most affected by group ability differences were the Levine equally and unequally reliable procedures. The equating procedures least affected by these ability differences were the Tucker and frequency estimation equating procedures. Matching on a set of common items, although pulling means obtained by the Tucker and frequency estimation equipercentile methods somewhat away from the criterion scaled-score mean, generally improved the scaled-score means obtained by the Levine and chained equipercentile procedures with respect to the criterion.

One might expect the Chemistry Test to behave in a somewhat similar manner to the Biology Test. Both tests are reflective of specific skills developed in one or perhaps two high school courses. The tests differ, however, in that chemistry is usually taught later in the high school curriculum than biology. Thus, examinees constituting spring and fall groups for the Chemistry Test may be closer together than Biology Test takers when recency of coursework is compared. This could cause the two tests to behave differently, which proved to be the case. For the Chemistry Test, the procedures most affected by differences in group ability were the Tucker and frequency estimation equipercentile procedures. The procedures least affected were the Levine equally and unequally reliable procedures. Matching on the basis of scores on a set of common items slightly worsened the Tucker and frequency estimation equipercentile results, when compared with the criterion scaled-score mean, and greatly worsened results obtained for the Levine and chained equipercentile procedures.

Scaled-score means for the American History and Social Studies Test, displayed in Figure 27, indicate that the equating procedure least affected by differences in group ability is the chained equipercentile procedure. The procedures most affected are the Levine equally and unequally reliable procedures. The Tucker and frequency estimation equipercentile procedures are affected by differences in group ability in a similar manner. Matching new- and old-form samples on the basis of common items was in this case quite unsuccessful, increasing the discrepancy between the experimental and criterion scaled-score means for all procedures except the Levine procedures.

Referring to the results displayed for the Mathematics Level II Test in Figure 24, one can see that the procedures most affected by differences in group ability are the Tucker and frequency estimation equipercentile procedures. The procedures least affected by differences in ability levels of the equating samples are the Levine equally and unequally reliable procedures. Matching on the basis of scores on common items improves the frequency estimation equipercentile equating and Tucker results slightly, when compared with the criterion scaled-score mean, but definitely worsens both the Levine and chained equipercentile results.

Finally, the results obtained for the French Test, which are displayed in Figure 33, indicate that the procedures most affected by differences in group ability are the Tucker and frequency estimation equipercentile methods. The procedures least affected by differences in group ability are the Levine equally and unequally reliable procedures. Matching on the basis of common items has almost no effect on the Tucker and frequency estimation equipercentile means but, as has been observed previously for the Chemistry and Mathematics Level II Tests, matching on common item scores worsens the results obtained for the Levine and chained equipercentile methods.

It is fairly clear from the analyses carried out for this study that using new- and old-form samples randomly selected from populations differing in ability level to equate Achievement Tests may lead to scores, and particularly scaled-score means, on the new forms of the tests that are less comparable than desired. In only a few instances, when the equating samples involved fall-spring random groups, the scaled-score means produced for particular new forms were similar to those obtained by the criterion equatings. These instances were (1) with the use of chained equipercentile equating for the American History and Social Studies Test, and (2) with the use of Tucker and frequency estimation equipercentile procedures with the Biology Test.

From an earlier study (Cook, Eignor, and Schmitt 1988), which examined experimental equatings for the Biology Test, it appeared that one solution to the problem of ability differences in equating samples was to match new- and old-form groups on some covariate such as a set of common items. Indeed, when this procedure was carried out for the Biology Test, the experimental equatings applied to the matched-group data produced scores consistent with each other and also closely approximating the criterion scores. Unfortunately, results observed for the matched (common items) equatings for the Biology Test did not generalize to the other four tests examined in this study. For the remaining tests studied, matching using common items produced close agreement among results, but these results were discrepant from the criterion results used for comparison.

Results of the experimental equatings for the Mathematics Level II Test indicate that if new- and old-form samples differ in level of ability, the equating methods that will produce scaled-score means closest to those obtained for the criterion equating are the Levine equally and unequally reliable methods. The results of the matched sample equatings carried out for the Mathematics Level II Test indicate that, for the Levine equally reliable, Levine unequally reliable, and chained equipercentile procedures, matching produces scaled-score means that are more discrepant from the criterion mean than not matching. The results of the frequency estimation and Tucker procedures used with the matched groups indicate that matching produces means slightly less discrepant from the criterion mean than if matching had not taken place. Generally, for the remaining tests, although one equating method may produce slightly better results (better in the sense that the scaled-score mean more closely matches the criterion mean), most methods are affected by differences in ability level between the new- and old-form samples. In all cases, however, except for the Levine equally and unequally reliable procedures used with the Biology and American History and Social Studies Tests, matching on a set of internal common items provides scaled-score means even more discrepant from the criterion means.

As mentioned previously, matching using responses to questions from the SDQ or the BQ was singularly unsuccessful. This is undoubtedly due to the low correlations between the respective covariates and Achievement Test scores. In addition, problems using SDQ responses to match groups in ability level were also probably related to the lower response rate to the SDQ questions used for this study.

One question that may be asked is: why did the Tucker and frequency estimation equating results for the Biology Test appear to be so invariant to differences in group ability? These two procedures when applied to the remaining four tests seemed to be quite affected by ability differences.

One of the main differences between the Biology Test and the remaining four tests studied is the length of the anchor test, 36 common items as opposed to a maximum number of 21 in the other tests. Perhaps because the additional number of items in the Biology Test common-item set provided a more reliable anchor-test score, scores on the anchor test contributed to the stability of the Tucker and frequency estimation equipercentile equating results across conditions. Some basis for this hypothesis exists in the literature.

Klein and Kolen (1985) investigated the relationship between accuracy of equating results and length of anchor test using the Tucker model. These researchers concluded that, "When the tests being equated were very similar, or in this particular case, identical, and the groups of examinees very similar, substantially more accurate equating was not obtained by lengthening the anchor. However, longer anchors did result in more accurate equating when the groups of examinees were dissimilar" (p. 10). An important question is whether or not the Tucker and frequency estimation equating results would have been more stable across the fall-spring random sampling combinations for the other four tests if the anchor test had been longer.

One confusing aspect of the Biology results is that the Biology Test common-item delta plot for the fall-spring random groups combination demonstrated the greatest scatter (see Figure 14) and the lowest correlation (.73) displayed by any of the five tests. One would expect that the poor relationship between the common items might have had a serious effect on all the equating procedures, not just the Levine equally and unequally reliable and the chained equipercentile procedures. The differential scatter in the equated delta plots observed for the fall-fall random, fall-spring random, and fall-spring matched (common items) samples contains important information for all the tests. What is most important to note is that only for the Biology Test, and to a much lesser extent the Chemistry Test, is the scatter more pronounced for the fall-spring random samples than for the fall-fall random samples. For the remaining three tests, the delta plots and accompanying correlation coefficients seem unaffected by differences in group ability. Also note that matching on distributions of scores on the set of common items had no effect on the scatter observed in the delta plots for any of the tests.

Angoff (1987) suggests that researchers should differentiate between groups that differ in ability level and groups that differ in patterns of performance on anchor-test items. He describes a situation in which, if groups were selected ". . . in such a way as to guarantee nearly equal overall means and standard deviations in the two groups, there is a likelihood that a significant interaction between item performance and group would still [be] found" (p. 293). He hypothesizes that this is due to the fact that the common items measure different psychological or educational traits for the two groups.

Following Angoff's reasoning, it could be possible that the fall-spring ability differences reflected in the common item scores do not have the same implications for all five tests. For the Biology Test, and to a lesser extent the Chemistry Test, these differences could indicate that the fall-spring groups differ not only in the ability measured by the tests, but in other fundamental educational traits. On the other hand, the groups taking the remaining three tests simply represent groups that are at different points on a linear continuum of ability.

Another interesting question is why matching on common items had such a positive effect on the Levine equally and unequally reliable results for the Biology Test and for the American History and Social Studies Test and such a negative effect on the Levine results for the other Achievement Tests.

Petersen, Cook, and Stocking (1983) compared equating results from Tucker, Levine equally reliable, Levine unequally reliable, and IRT procedures applied in an anchor-test design used to equate the SAT. These researchers concluded that, of the conventional procedures evaluated, the

most stable equating results were provided by the Levine equally reliable model, and the least stable results were provided by the Tucker model. Petersen, et al., explained their results as follows: "Implicit to the derivation of the Tucker model is the assumption of random groups (Angoff 1984; Levine 1955). Because the samples for the test editions to be equated were not random samples from the same administration, and in several instances, differed considerably in ability level, it is not surprising that the Levine models gave more satisfactory results than the Tucker model" (p. 152). The explanation provided by Petersen, et al., is appropriate as an explanation of the Tucker and Levine equally reliable results obtained in the current study for the Mathematics Level II, Chemistry, and French Tests. It does not, however, provide an explanation for the results obtained for the Biology and American History and Social Studies Tests.

The results obtained in the current study for the Chemistry Test were somewhat perplexing. It was expected that equating results obtained for this test would be similar to those obtained for the Biology Test; however, this similarity was not observed. For the Chemistry Test, the Tucker and frequency estimation equating scaled-score means observed for the fall-spring random samples were very different from those observed for the criterion equating. What makes this observation even more surprising is that the differences in ability level observed for the Chemistry Test fall-spring random groups (as evidenced by scores on the common-item set) were the smallest of any of the five tests included in this study.

Conclusions

The results of the current study should be considered tentative for several reasons. For one, the criterion used to evaluate the experimental equatings was a very pragmatic one. The criterion for each test in the study was based on the sampling plan that is currently used to provide reported scores for the particular Achievement Test investigated. The question explored was: what would be the effect on reported scores of changing the currently used sampling plan? And, given that a serious effect was observed, could this effect be ameliorated by matching new- and old-form samples on ability level using distributions of scores on a set of common items or on some other covariate, such as responses to a background questionnaire given at the time of testing or when an examinee first registers to take a test? Use of a different equating criterion might possibly have led to a different interpretation of the results of the study. However, the chosen criterion has practical value for the specific questions investigated. A second limitation of this study is a lack of replication. Had different test forms or samples been chosen for this study, the results of the study might have differed.

In spite of the caveats mentioned above, several conclusions can be drawn from the results of this study. The most important conclusion appears to be that it may be difficult, and in some cases impossible, to equate Achievement Tests using new- and old-form samples obtained from different populations of examinees. It appears as though all equating methods are affected by these differences to some extent. In general, the equating models that appear to be the most affected by differences in group ability are the Tucker and frequency estimation equipercentile models. The models that appear to be the most robust to differences in group ability are the chained equipercentile and, in some instances, the Levine models.

Matching on the basis of observed scores on a set of internal common items does not remedy the situation. In general, matching produces results for all equating procedures that are similar but that over- or underestimate scaled scores. Because the results are similar across methods, the effect can be quite misleading in that, in the absence of a criterion, one could conclude that because consistent results are obtained across methods, the results are close to "truth." This was not found to be the case for the situations investigated in this study.

To summarize, the results of the study cast serious doubt on the viability of an anchor-test data collection design when samples that differ in ability are used to equate Achievement Tests of the type investigated in this study. It appears as though the equating models evaluated in this study are seriously affected by differences in group ability, and that this effect occurs even when there is little group-by-item interaction displayed by the common-item set. Further, attempts to ameliorate the effect of group differences by matching on a covariate such as the common-item set itself were quite unsuccessful, and matching is not recommended as a procedure for rectifying the problem of group ability-level differences when using an anchor-test equating design to equate Achievement Tests such as those used in this study.

An important consideration for future study is what, if any, effect does equating tests based on new- and old-form samples selected randomly from spring administrations (as compared with fall administrations) have on equated scores. The question remains whether or not it would be a prudent and appropriate policy to introduce Achievement Tests at multiple time points in the school year as long as new- and old-form samples are both randomly selected from, say, two December or two June administrations. It is possible that introduction of new forms should be restricted to only fall or only spring administrations. Further study to address this question is recommended.

REFERENCES

Angoff, W. H. 1984. *Scales, norms, and equivalent scores.* Princeton, N.J.: Educational Testing Service.

Angoff, W. H. 1987. Technical and practical issues in equating: A discussion of four papers. *Applied Psychological Measurement* 11: 291–300.

Cook, L. L., D. R. Eignor, and A. P. Schmitt. 1988. *The effects on IRT and conventional achievement test equating results of using equating samples matched on ability* (RR-88-52). Princeton, N.J.: Educational Testing Service.

Cook, L. L., D. R. Eignor, and H. L. Taft. 1988. A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement* 25: 31–45.

Cureton, E. E., and J. W. Tukey. 1951. Smoothing frequency distributions, equating tests, and preparing norms. *American Psychologist* 6:404.

Fairbank, B. A. 1987. The use of presmoothing and postsmoothing to increase the precision of equipercentile equating. *Applied Psychological Measurement* 11: 245–262.

Hecht, L. W., and F. Swineford. 1981. *Item analysis at Educational Testing Service*. Princeton, N.J.: Educational Testing Service.

Henrysson, S. 1971. Gathering, analyzing, and using data on test items. In *Educational measurement* (2nd ed.), edited by R. L. Thorndike. Washington, D.C.: American Council on Education.

Holland, P. W., and D. B. Rubin. 1982. *Test equating*. New York: Academic Press.

Klein, L. W., and M. J. Kolen. 1985. Effect of number of common items in common-item equating with nonrandom groups. Paper presented at the annual meeting of American Educational Research Association, Chicago, April.

Lawrence, I. L., and N. J. Dorans. 1988. *A comparison of observed score and true score equating methods for representative samples and samples matched on an anchor test* (RR-88-23). Princeton, N.J.: Educational Testing Service.

Levine, R. S. 1955. *Equating the score scales of alternate forms administered to samples of different ability* (ETS Research Bulletin No. 23). Princeton, N.J.: Educational Testing Service.

Lord, F. M. 1975. Automated hypothesis tests and standard errors for nonstandard problems. *The American Statistician* 29: 56–59.

Petersen, N. S., L. L. Cook, and M. L. Stocking. 1983. IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics* 8: 137–156.

Rosenbaum, P. R., and D. T. Thayer. 1986. *Smoothing the joint and marginal distributions of scored two-way contingency tables in test equating* (Research Statistics Report). Princeton, N.J.: Educational Testing Service.

Rosenbaum, P. R., and D. T. Thayer. 1987. Smoothing the joint and marginal distributions of scored two-way contingency tables in test equating. *British Journal of Mathematical and Statistical Psychology* 40: 43–49.

Score	Fall new form group		Spring old form total group		Spring old form matched group	
	Frequency	Percent Below	Frequency	Percent Below	Frequency	Percent Below
36	6	99.9	133	99.4	6	99.9
35	10	99.3	342	97.9	10	99.3
34	19	98.5	507	95.7	19	98.5
33	3	98.4	184	94.9	3	98.4
32	26	97.3	644	92.1	26	97.3
31	40	95.7	871	88.4	40	95.7
30	47	93.7	1026	84.0	47	93.7
29	70	90.8	1229	78.7	70	90.8
28	20	90.0	591	76.2	20	90.0
27	64	87.3	1078	71.6	64	87.3
26	67	84.6	1297	66.0	67	84.6
25	101	80.4	1348	60.2	101	80.4
24	111	75.7	1389	54.3	111	75.7
23	68	72.9	760	51.0	68	72.9
22	96	68.9	1118	46.2	96	68.9
21	99	64.8	1178	41.2	99	64.8
20	120	59.8	1124	36.4	120	59.8
19	116	55.0	1160	31.4	116	55.0
18	85	51.5	603	28.8	85	51.5
17	131	46.1	920	24.9	131	46.1
16	125	40.9	888	21.1	125	40.9
15	110	36.3	890	17.3	110	36.3
14	133	30.8	758	14.1	133	30.8
13	81	27.4	389	12.4	81	27.4
12	89	23.7	567	10.0	89	23.7
11	99	19.8	534	7.7	99	19.6
10	100	15.4	456	5.7	100	15.4
9	85	11.9	373	4.1	85	11.9
8	56	9.6	175	3.4	56	9.6
7	68	6.8	232	2.4	68	6.8
6	42	5.0	190	1.6	42	5.0
5	37	3.5	124	1.1	37	3.5
4	28	2.3	124	0.6	28	2.3
3	13	1.8	41	0.4	13	1.8
2	19	1.0	50	0.2	19	1.0
1	12	0.5	27	0.1	12	0.5
0	6	0.2	29	0.1	6	0.2
-1	5	0.0	13	0.0	5	0.0
-2	0	0.0	4	0.0	0	0.0
-3	0	0.0	1	0.0	0	0.0
-4	1	0.0	1	0.0	1	0.0
-5	0	0.0	1	0.0	0	0.0
	2,408		23,369		2,408	

Figure 1. Distributions of scores and percentages below on the 36 common items used in matching the Biology Test spring old-form group to the fall new-form group.

For each of the subject areas in questions 12 through 17, blacken the *latest* year-end or midyear grade you received since beginning the ninth grade. For example, if you are a senior and have not taken biology or any other biological science since your sophomore year, indicate that year-end grade. If you are a junior and have completed the first half of the year in an English course, indicate that midyear grade.

If you received the grade in an advanced, accelerated, or honors course, also blacken the letter H.

(A) Excellent (usually 90-100 or A)
(B) Good (usually 80-89 or B)
(C) Fair (usually 70-79 or C)
(D) Passing (usually 60-69 or D)
(F) Failing (usually 59 or below or F)
(G) Only "pass-fail" grades were assigned and I received a pass.
(H) The grade reported was in an advanced, accelerated, or honors course.

12. English

13. Mathematics

14. Foreign Languages

15. Biological Sciences

16. Physical Sciences

17. Social Studies

Original Response	Recoded Numerical Response
F	1
D	2
C&G	3
B	4
A	5
H	If H is marked, advance students recoded numerical response by one

Correlation of question 15 responses with Biology Test scores	Fall new form	Spring old form
	.42	.45

Figure 2. Biology SDQ question 15 and specifications for the recoding of responses for cross-tabulation and matching purposes; correlations of recoded responses with Biology Achievement Test scores for fall new-form and spring old-form groups.

Questions 47 through 60 concern how you feel you compare with other people your own age in certain areas of ability. For each field, blacken the letter

(A) if you feel you are in the highest 1 percent in that area of ability
(B) if you feel you are in the highest 10 percent in that area of ability
(C) if you feel you are above average in that area of ability
(D) if you feel you are average in that area of ability
(E) if you feel you are below average in that area of ability

47. Acting ability
48. Artistic ability
49. Athletic ability
50. Creative writing
51. Getting along with others
52. Leadership ability
53. Mathematical ability
54. Mechanical ability
55. Musical ability
56. Organizing work
57. Sales ability
58. Scientific ability
59. Spoken expression
60. Written expression

Original Response	Recoded Numerical Response
E	1
D	2
C	3
B	4
A	5

Correlation of question 58 responses with Biology Test scores	Fall new form	Spring old form
	.40	.43

Figure 3. Biology SDQ question 58 and specifications for the recoding of responses for cross-tabulation and matching purposes; correlations of recoded responses with Biology Achievement Test scores for fall new-form and spring old-form groups.

- Biology -

Recoded Question 58

Recoded Question 15	1	2	3	4	5
1	0	0	0	0	0
2	0	0	1	3	1
3	0	24	18	5	4
4	6	67	132	104	21
5	0	59	197	295	166
6	0	7	35	113	90

Fall new form group

1348
frequencies missing = 1058
percent missing = 44%

Recoded Question 58

Recoded Question 15	1	2	3	4	5
1	0	3	0	0	0
2	8	29	9	3	1
3	28	299	180	57	12
4	43	752	1247	758	163
5	10	523	1577	2153	1011
6	5	112	538	1233	1027

Spring old form total group

11,826
frequencies missing = 11,543
percent missing = 49%

Recoded Question 58

Recoded Question 15	1	2	3	4	5
1	0	0	0	0	0
2	0	0	1	3	1
3	0	24	18	5	4
4	6	67	132	104	21
5	0	59	197	295	166
6	0	7	35	113	90

Spring old form matched group

1348

Figure 4. Cross-tabulations of recoded responses to SDQ questions 15 and 58 used in matching the Biology spring old-form group to the fall new-form group.

- Mathematics Level II -

Score	Fall new form group		Spring old form total group		Spring old form matched group	
	Frequency	Percent Below	Frequency	Percent Below	Frequency	Percent Below
19	24	99.0	42	97.9	15	99.0
18	42	97.1	95	93.2	26	97.2
17	44	95.2	106	87.9	28	95.2
16	58	92.7	86	83.6	37	92.7
15	101	88.3	139	76.7	64	88.2
14	131	82.6	145	69.5	83	82.5
13	182	74.7	159	61.6	115	74.5
12	196	66.2	184	52.4	123	66.0
11	186	58.2	167	44.1	117	57.9
10	226	48.4	173	35.5	142	48.0
9	213	39.1	148	28.1	134	38.7
8	201	30.4	133	21.5	127	29.9
7	162	23.3	104	16.3	104	22.7
6	135	17.5	84	12.1	84	16.9
5	122	12.2	76	8.4	76	11.6
4	90	8.3	57	5.5	57	7.7
3	67	5.4	34	3.8	34	5.3
2	54	3.0	37	2.0	37	2.8
1	22	2.1	18	1.1	18	1.5
0	30	0.8	13	0.4	13	0.6
-1	14	0.2	3	0.3	3	0.4
-2	3	0.0	4	0.1	4	0.1
-3	1	0.0	1	0.0	1	0.0
-4	0	0.0	1	0.0	1	0.0
	2,304		2,009		1,443	

Figure 5. Distributions of scores and percentages below on the 19 common items used in matching the Mathematics Level II Test spring old-form group to the fall new-form group.

- American History -

Score	Fall new form group		Spring old form total group		Spring old form matched group	
	Frequency	Percent Below	Frequency	Percent Below	Frequency	Percent Below
20	3	99.9	20	99.0	2	99.9
19	6	99.6	43	96.9	4	99.5
18	21	98.6	68	93.5	14	98.5
17	15	97.8	37	91.7	10	97.7
16	49	95.5	96	87.0	33	95.3
15	71	92.1	112	81.5	47	91.7
14	121	86.2	171	73.1	81	85.6
13	119	80.5	145	65.9	79	79.7
12	97	75.8	142	58.9	65	74.8
11	157	68.3	161	51.0	105	66.9
10	152	61.0	181	42.1	101	59.3
9	197	51.5	172	33.6	131	49.4
8	204	41.7	146	26.4	136	39.2
7	162	33.9	108	21.1	108	31.1
6	159	26.2	120	15.2	106	23.1
5	170	18.0	103	10.1	103	15.4
4	133	11.6	76	6.4	76	9.6
3	106	6.5	46	4.1	46	6.2
2	64	3.5	35	2.4	35	3.5
1	35	1.8	24	1.2	23	1.8
0	22	0.7	14	0.5	14	0.8
-1	6	0.4	7	0.2	5	0.3
-2	6	0.1	2	0.1	2	0.2
-3	1	0.1	1	0.0	1	0.1
-4	2	0.0	1	0.0	1	0.0
	2,078		2,031		1,329	

Figure 6. Distributions of scores and percentages below on the 20 common items used in matching the American History and Social Studies Test spring old-form group to the fall new-form group.

- Chemistry -

Score	Fall new form group		Spring old form total group		Spring old form matched group	
	Frequency	Percent Below	Frequency	Percent Below	Frequency	Percent Below
20	13	99.4	22	99.0	9	99.4
19	44	97.2	69	95.9	32	97.1
18	56	94.4	111	90.8	40	94.4
17	43	92.3	40	89.0	31	92.2
16	96	87.5	115	83.8	69	87.4
15	126	81.3	167	76.2	91	81.1
14	131	74.8	149	69.5	94	74.5
13	155	67.1	192	60.8	112	66.7
12	128	60.7	147	54.1	92	60.3
11	141	53.7	202	45.0	102	53.2
10	194	44.1	202	35.8	140	43.5
9	182	35.1	162	28.5	131	34.3
8	138	26.3	182	20.2	99	27.4
7	112	22.7	113	15.1	81	21.8
6	125	16.5	101	10.5	90	15.5
5	119	10.6	80	6.9	80	10.0
4	88	6.2	63	4.0	63	5.6
3	59	3.3	41	2.2	41	2.7
2	30	1.8	18	1.4	18	1.5
1	15	1.1	20	0.5	11	0.7
0	16	0.3	7	0.1	7	0.2
-1	5	0.1	3	0.0	3	0.0
-2	1	0.0	0	0.0	0	0.0
	2,017		2,206		1,436	

Figure 7. Distributions of scores and percentages below on the 20 common items used in matching the Chemistry Test spring old-form group to the fall new-form group.

- French -

Score	Fall new form group		Spring old form total group		Spring old form matched group	
	Frequency	Percent Below	Frequency	Percent Below	Frequency	Percent Below
21	14	99.8	39	99.4	7	99.8
20	44	99.1	93	97.8	23	99.1
19	3	99.0	19	97.6	2	99.0
18	84	97.6	186	94.6	45	97.6
17	172	94.8	252	90.4	91	94.8
16	253	90.7	333	85.0	134	90.7
15	71	89.5	86	83.5	38	89.5
14	303	84.6	343	77.9	161	84.6
13	420	77.7	450	70.5	223	77.7
12	451	70.4	514	62.0	239	70.4
11	195	67.2	207	58.6	103	67.2
10	484	59.3	486	50.6	257	59.3
9	528	50.7	528	42.0	280	50.6
8	609	40.7	542	33.0	323	40.7
7	328	35.4	249	28.9	174	35.3
6	478	27.6	395	22.4	253	27.6
5	469	19.9	404	15.8	249	19.9
4	398	13.4	315	10.6	211	13.4
3	208	10.0	145	8.2	110	10.0
2	230	6.3	212	4.7	122	6.3
1	172	3.4	155	2.2	91	3.4
0	114	1.6	60	1.2	60	1.6
-1	40	0.9	27	0.8	21	1.0
-2	39	0.3	26	0.3	21	0.3
-3	11	0.1	13	0.1	6	0.1
-4	7	0.0	5	0.0	4	0.0
-5	0	0.0	2	0.0	0	0.0
	6,125		6,086		3,248	

Figure 8. Distributions of scores and percentages below on the 21 common items used in matching the French Test spring old-form group to the fall new-form group.

Questions 6 through 11 ask you to blacken the letter corresponding to the total years of study you expect to complete in certain subject areas. Include in the total only courses you have taken since beginning the ninth grade and those you expect to complete before graduation from high school. Count less than a full year in a subject as a full year. Do not count a repeated year of the same course as an additional year of study.

(A) One year or the equivalent
(B) Two years or the equivalent
(C) Three years or the equivalent
(D) Four years or the equivalent
(E) More than four years or the equivalent
(F) I will not take any courses in the subject area.

6. English
7. Mathematics
8. Foreign Languages
9. Biological Sciences (for example, biology, botany, or zoology)
10. Physical Sciences (for example, chemistry, physics, or earth science)
11. Social Studies (for example, history, government, or geography)

Original Response	Recoded Numerical Response
F	1
A	2
B	3
C	4
D	5
E	6

Correlation of question 8 responses with French Test scores	Fall new form	Spring old form
	.30	.32

Figure 9. French SDQ question 8 and specifications for the recoding of responses for cross-tabulation and matching purposes; correlations of recoded responses with French Achievement Test scores for fall new-form and spring old-form groups.

For each of the subject areas in questions 12 through 17, blacken the *latest* year-end or midyear grade you received since beginning the ninth grade. For example, if you are a senior and have not taken biology or any other biological science since your sophomore year, indicate that year-end grade. If you are a junior and have completed the first half of the year in an English course, indicate that midyear grade.

If you received the grade in an advanced, accelerated, or honors course, also blacken the letter H.

(A) Excellent (usually 90-100 or A)
(B) Good (usually 80-89 or B)
(C) Fair (usually 70-79 or C)
(D) Passing (usually 60-69 or D)
(F) Failing (usually 59 or below or F)
(G) Only "pass-fail" grades were assigned and I received a pass.
(H) The grade reported was in an advanced, accelerated, or honors course.

12. English

13. Mathematics

14. Foreign Languages

15. Biological Sciences

16. Physical Sciences

17. Social Studies

Original Response	Recoded Numerical Response
F	1
D	2
C&G	3
B	4
A	5
H	If H is marked, advance students recoded numerical response by one

	Fall new form	Spring old form
Correlation of question 14 responses with French Test scores	.27	.38

Figure 10. French SDQ question 14 and specifications for the recoding of responses for cross-tabulation and matching purposes; correlations of recoded responses with French Achievement Test scores for fall new-form and spring old-form groups.

Recoded Question 14

Recoded Question 8	1	2	3	4	5	6
1	0	0	0	2	2	1
2	0	0	0	2	6	3
3	0	1	3	21	32	2
4	0	0	25	169	247	80
5	0	1	55	404	820	280
6	0	1	12	154	450	208

2,961

Fall new form group

frequencies missing = 3,164
percent missing = 51.7

Recoded Question 14

Recoded Question 8	1	2	3	4	5	6
1	0	0	0	3	0	1
2	0	0	2	2	12	3
3	0	2	17	33	56	8
4	0	11	68	295	362	114
5	0	5	89	559	1032	507
6	0	2	30	195	504	374

4,286

Spring old form total group

frequencies missing = 1,792
percent missing = 29.5

Recoded Question 14

Recoded Question 8	1	2	3	4	5	6
1	0	0	0	2	0	1
2	0	0	0	2	6	3
3	0	1	3	21	32	2
4	0	0	25	169	247	80
5	0	1	55	404	820	280
6	0	1	12	154	450	208

2,959

Spring old form matched group

Figure 11. Cross-tabulations of recoded responses to SDQ questions 8 and 14 used in matching the French spring old-form group to the fall new-form group.

In the group of nine ovals labeled Q, you are to fill in ONE and ONLY ONE oval, as described below, to indicate how you obtained your knowledge of French. The information that you provide is for statistical purposes only and will not influence your score on the test.

If your knowledge of French does not come primarily from courses taken in grades 9 through 12, fill in oval 9 and leave the remaining ovals blank, regardless of how long you studied the subject in school. For example, you are to fill in oval 9 if your knowledge of French comes primarily from any of the following sources: study prior to the ninth grade, courses taken at a college, special study, residence abroad, or living in a home in which French is spoken.

If your knowledge of French does come primarily from courses taken in grades 9 through 12, fill in the oval that indicates the level of the French course in which you are currently enrolled. If you are not now enrolled in a French course, fill in the oval that indicates the level of the most advanced course in French that you have completed.

Level I:	first or second half	- fill in oval 1
Level II:	first half	- fill in oval 2
	second half	- fill in oval 3
Level III:	first half	- fill in oval 4
	second half	- fill in oval 5
Level IV:	first half	- fill in oval 6
	second half	- fill in oval 7
Advanced Placement or course that represents a level of study higher than Level IV: second half		- fill in oval 8

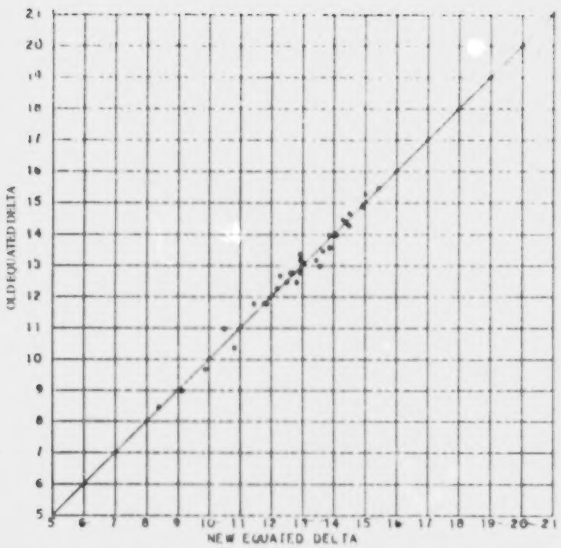
If you are in doubt about whether to mark oval 9 rather than one of the ovals 1-8, mark oval 9.

Figure 12. Question on Background Questionnaire in French used in matching the French spring old-form group to the fall new-form group.

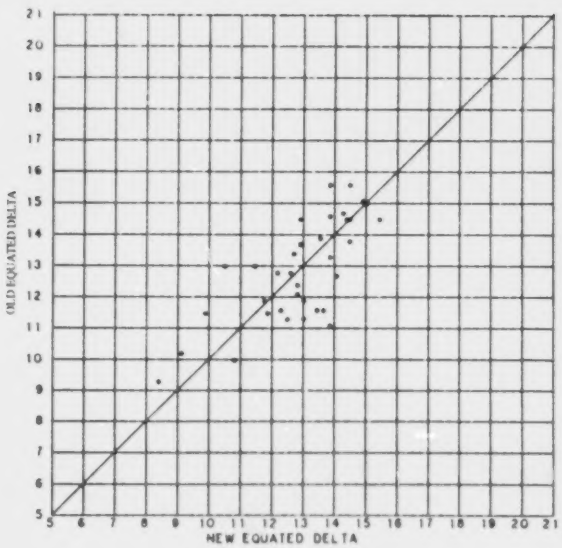
- French -

Oval filled	Fall new form group		Spring old form total group		Spring old form matched group	
	Frequency	Percent Below	Frequency	Percent Below	Frequency	Percent Below
9	0	100.0	0	100.0	0	100.0
8	1632	73.4	550	91.0	247	73.4
7	731	61.4	2283	53.4	111	61.4
6	2271	24.3	344	47.7	344	24.3
5	710	12.8	2242	10.8	108	12.7
4	596	3.0	298	5.9	90	3.0
3	185	0.0	361	0.0	28	0.0
2	0	0.0	0	0.0	0	0.0
1	0	0.0	0	0.0	0	
	6,125		6,078		928	

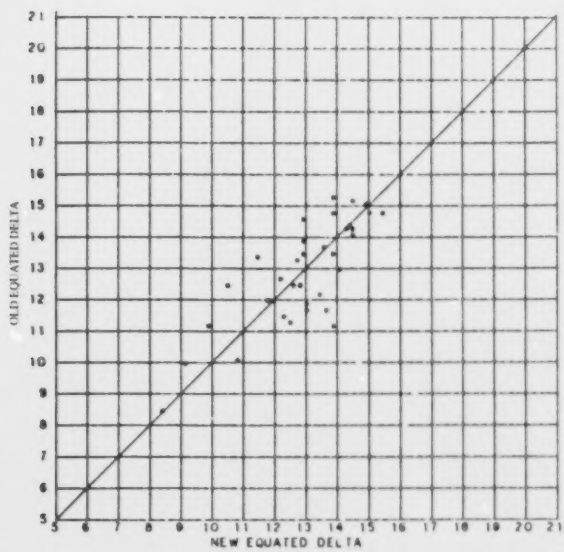
Figure 13. Distributions of responses to question on Background Questionnaire in French used in matching the French spring old-form group to the fall new-form group.



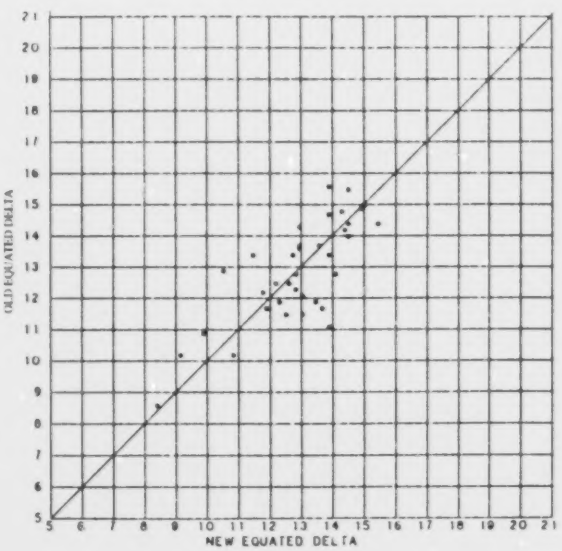
r = .99
fall/fall random groups



r = .73
fall/spring random groups

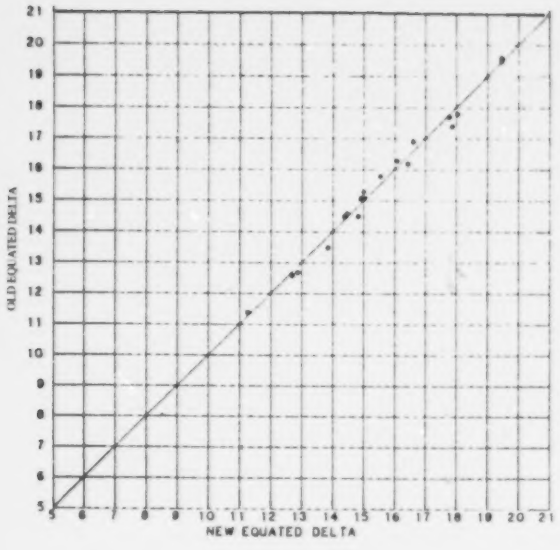


r = .79
fall/spring matched groups
(common items)

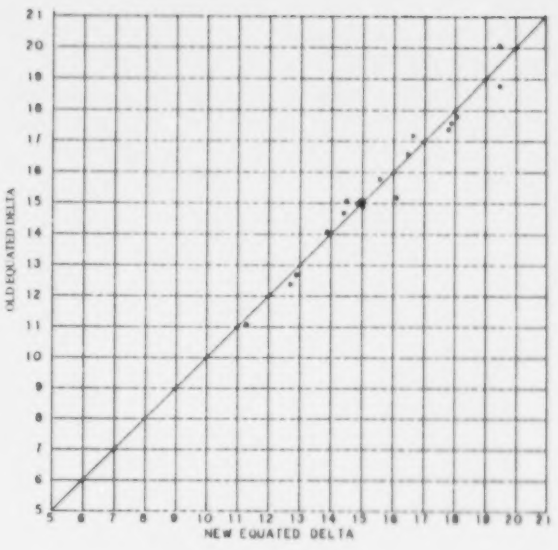


r = .77
fall/spring matched groups
(SDQ)

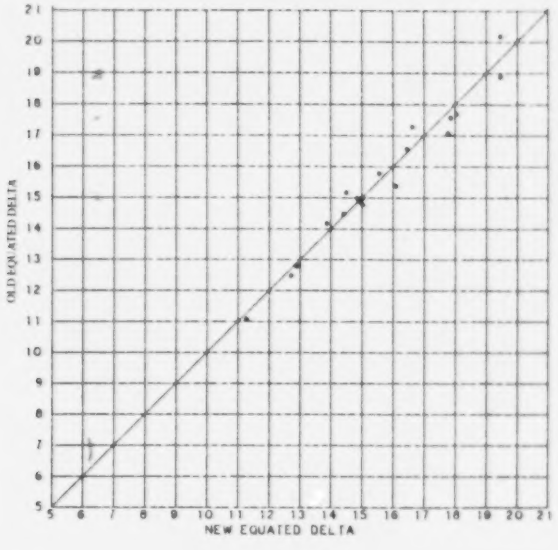
Figure 14. Plots of Biology fall new-form versus fall old-form equated deltas for the one fall-fall sample combination and versus spring old-form equated deltas for the three fall-spring sample combinations.



r = .99
fall/fall random groups

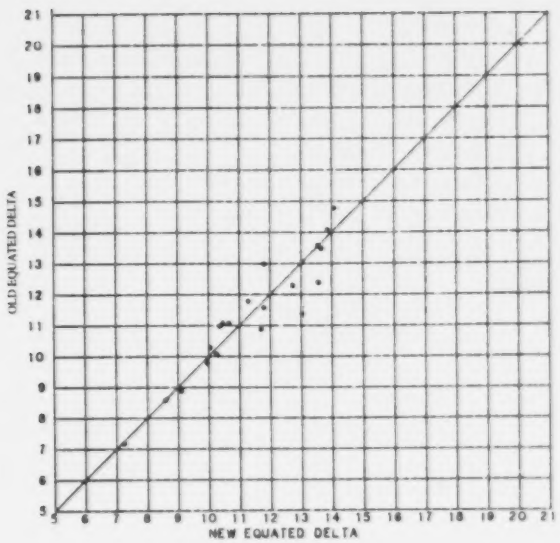


r = .98
fall/spring random groups

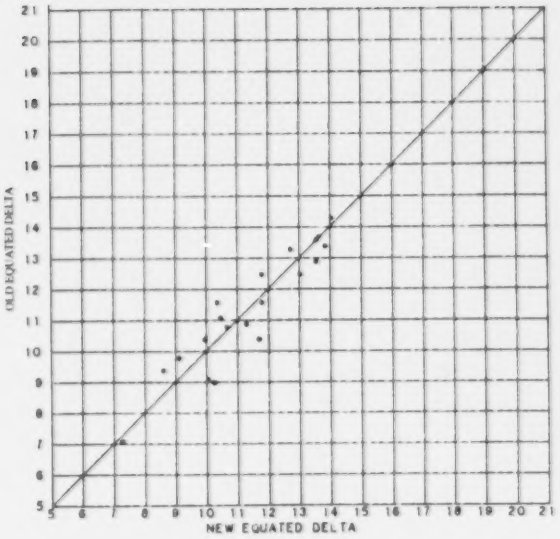


r = .98
fall/spring matched groups
(common items)

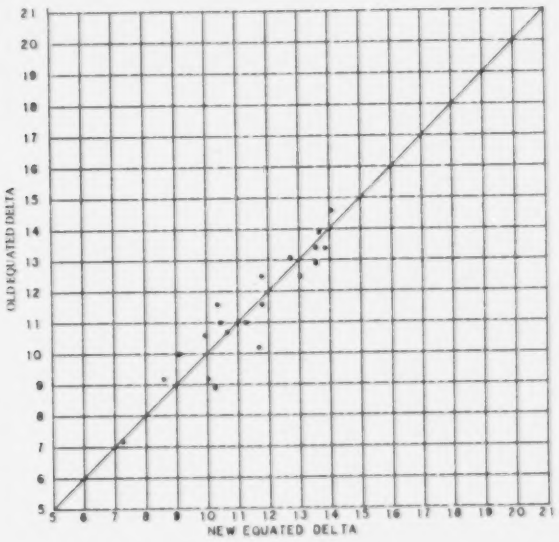
Figure 15. Plots of Mathematics Level II fall new-form versus fall old-form equated deltas for the one fall-fall sample combination and versus spring old-form equated deltas for the two fall-spring sample combinations.



r = .94
fall/fall random groups

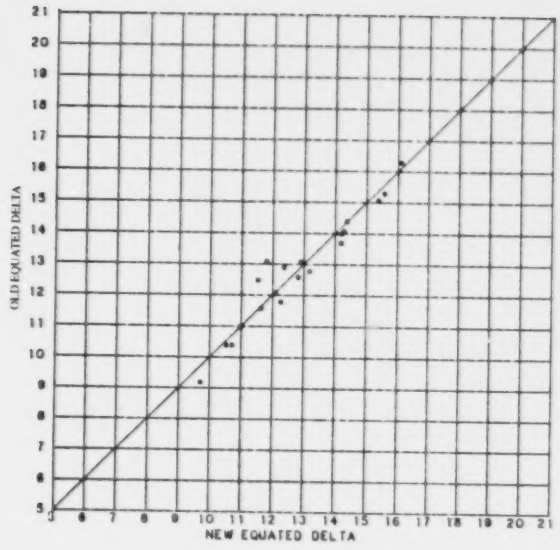


r = .93
fall/spring random groups

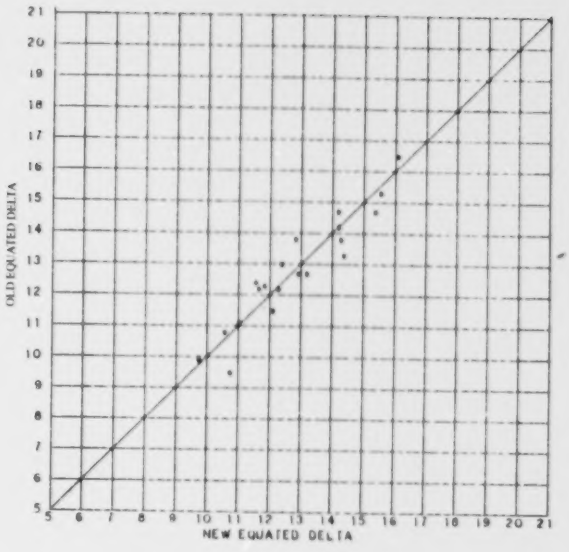


r = .93
fall/spring matched groups
(common items)

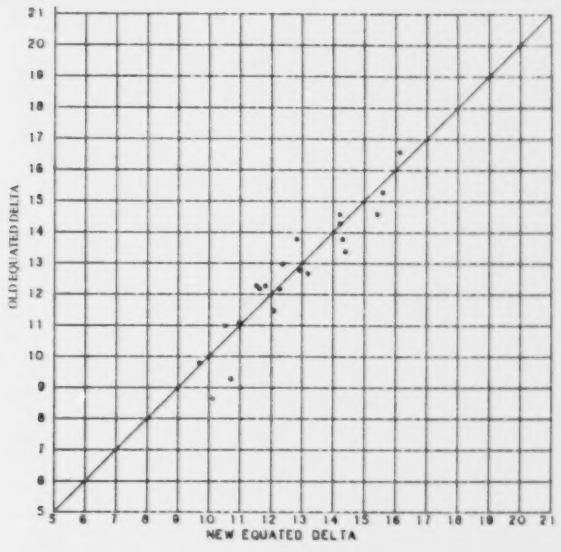
Figure 16. Plots of American History and Social Studies fall new-form versus fall old-form equated deltas for the one fall-fall sample combination and versus spring old-form equated deltas for the two fall-spring sample combinations.



r = .97
fall/fall random groups

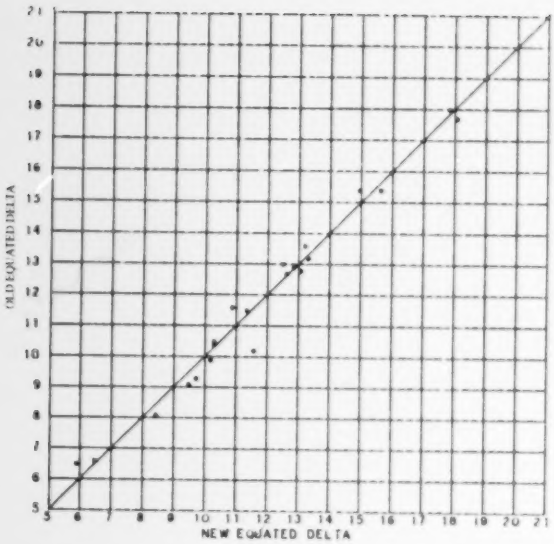


r = .94
fall/spring random groups

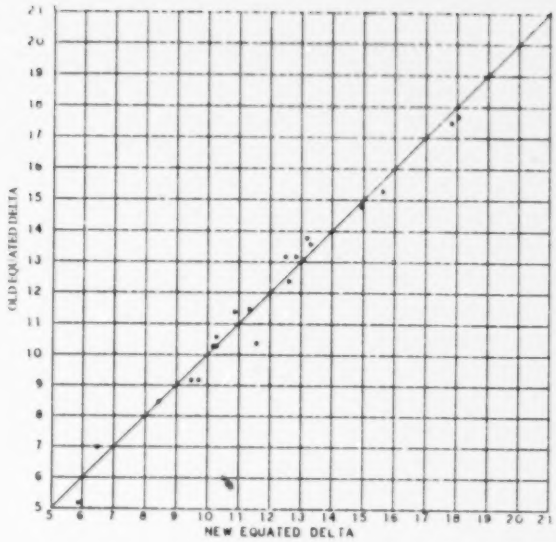


r = .94
fall/spring matched groups
(common items)

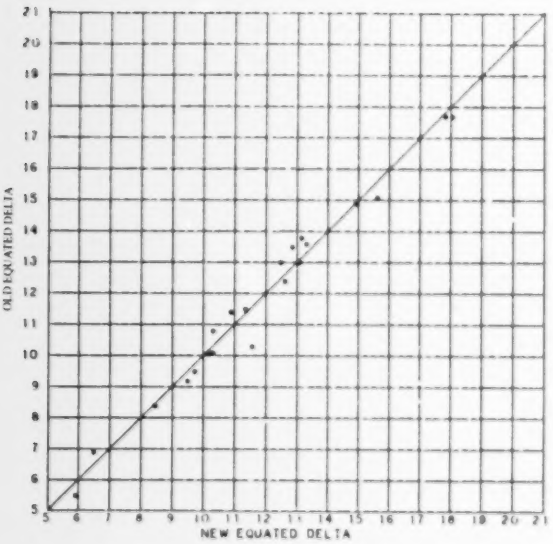
Figure 17. Plots of Chemistry fall new-form versus fall old-form equated deltas for the one fall-fall sample combination and versus spring old-form equated deltas for the two fall-spring sample combinations.



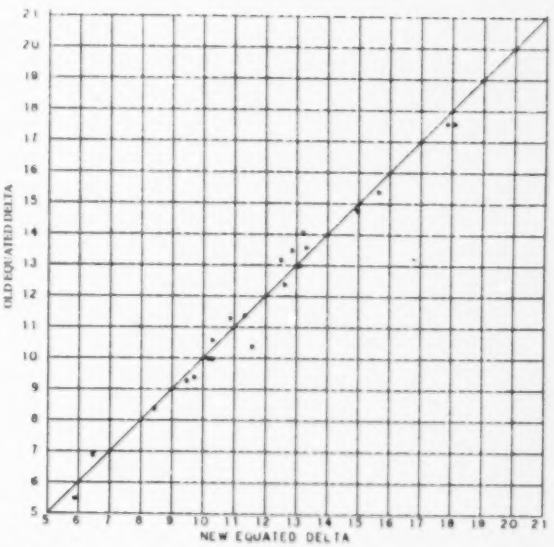
r = .99
fall/fall random groups



r = .99
fall/spring random groups

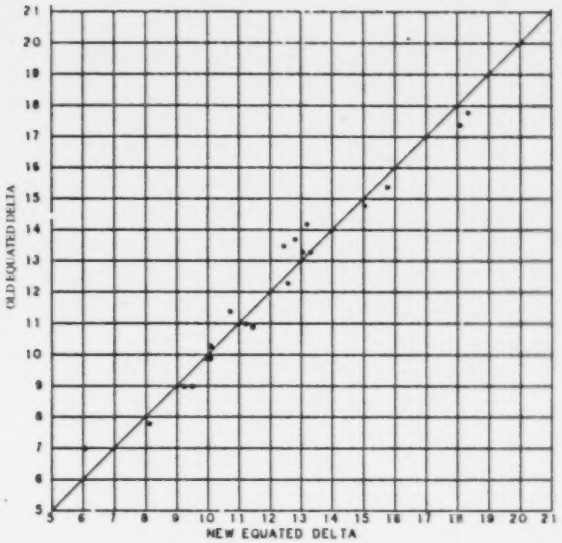


r = .99
fall/spring matched groups
(common items)

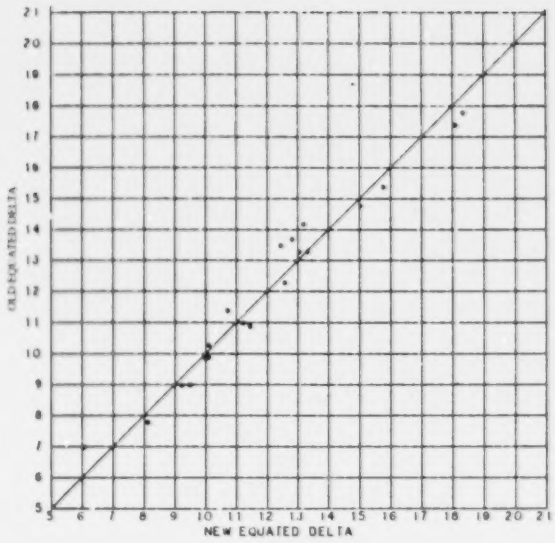


r = .99
fall/spring matched groups
(SDO)

Figure 18. Plots of French fall new-form versus fall old-form equated deltas for the one fall-fall sample combination and versus spring old-form equated deltas for the four fall-spring sample combinations.

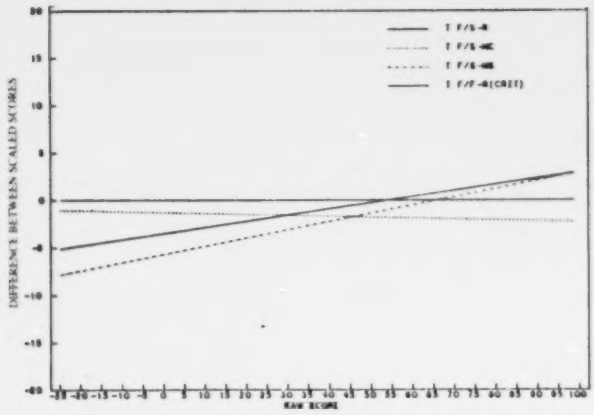


r = .98
fall/spring matched groups
(BO)



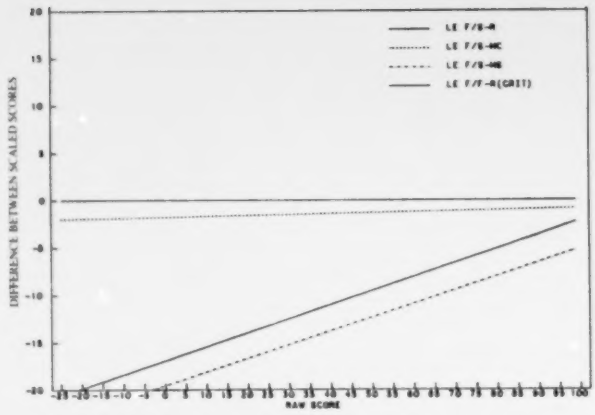
r = .98
fall/spring matched groups
(BO)

Figure 18. (continued)



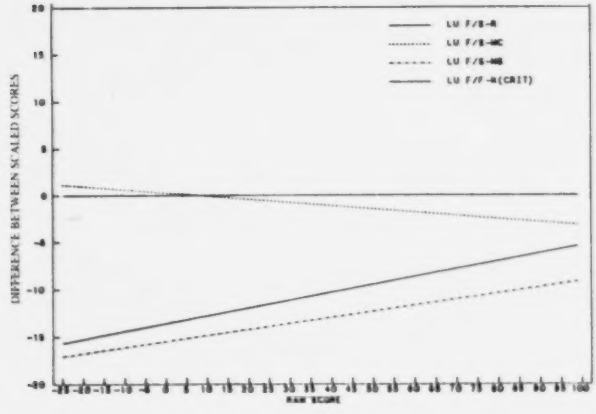
(a)

fall-spring random or matched groups Tucker equatings—fall-fall random groups Tucker criterion equating



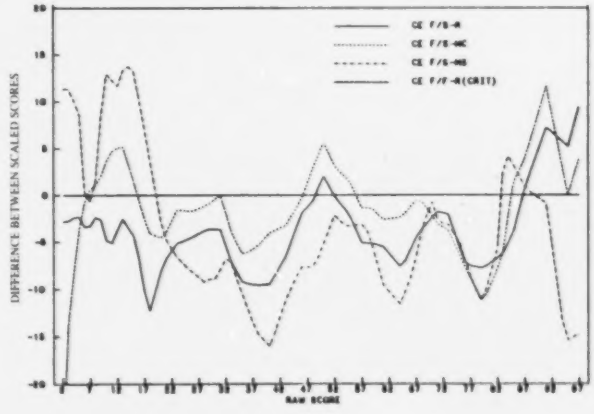
(b)

fall-spring random or matched groups Levine equally reliable equatings—fall-fall random groups Levine equally reliable criterion equating



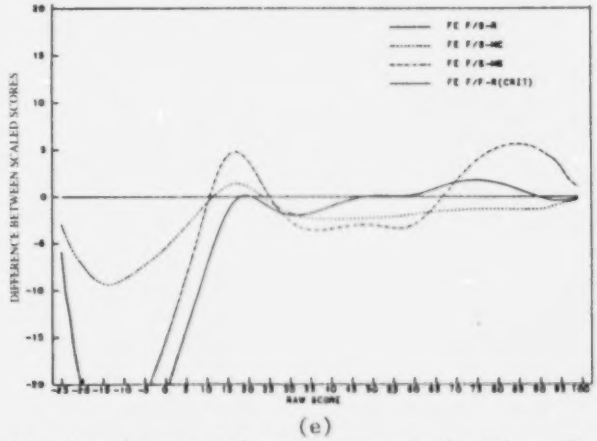
(c)

fall-spring random or matched groups Levine unequally reliable equatings—fall-fall random groups Levine unequally reliable criterion equating



(d)

fall-spring random or matched groups chained equipercentile equatings—fall-fall random groups chained equipercentile criterion equating

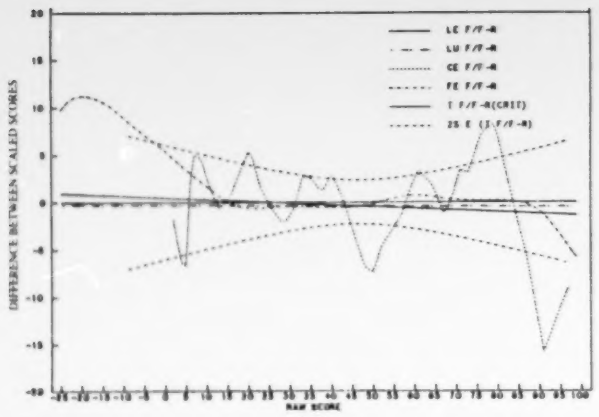


(e)

fall-spring random or matched groups frequency estimation equipercentile equatings—fall-fall random groups frequency estimation equipercentile criterion equating

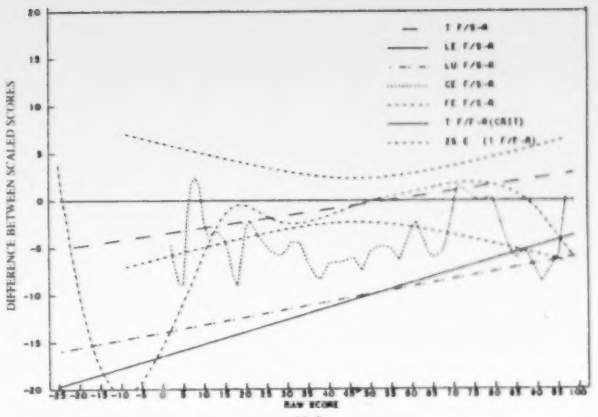
Abbreviations used in plots: fall-fall random groups: F/F-R; fall-spring random groups: F/S-R; fall-spring matched groups (common items): F/S-MC; fall-spring matched groups (SDQ): F/S-MS; Tucker: T; Levine equally reliable: LE; Levine unequally reliable: LU; chained equipercentile: CE; frequency estimation equipercentile: FE

Figure 19. Biology equating-difference plots (fall-spring random or matched groups equatings minus fall-fall random groups criterion equating).



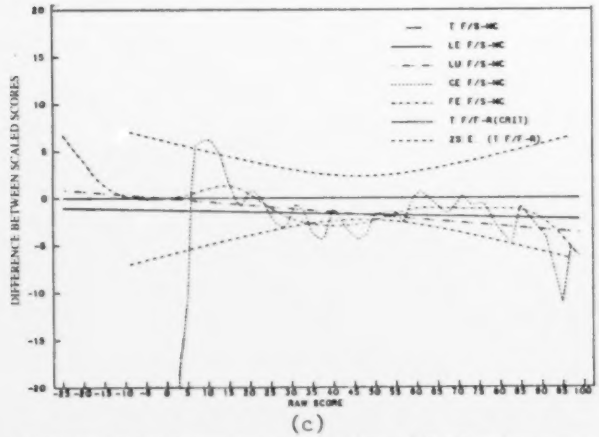
(a)

other fall-fall random groups equatings—fall-fall random groups Tucker criterion equating



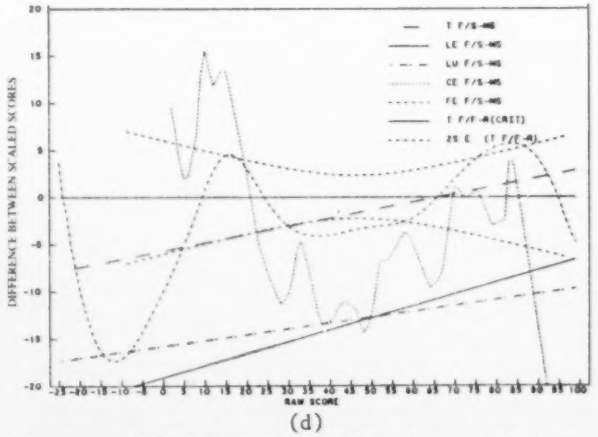
(b)

fall-spring random groups equatings—fall-fall random groups Tucker criterion equating



(c)

fall-spring matched (common items) equatings—fall-fall random groups Tucker criterion equating



(d)

fall-spring matched (SDQ) equatings—fall-fall random groups Tucker criterion equating

Abbreviations used in plots: fall-fall random groups: F/F-R; fall-spring random groups: F/S-R; fall-spring matched groups (common items): F/S-MC; fall-spring matched groups (SDQ): F/S-MS; Tucker: T; Levine equally reliable: LE; Levine unequally reliable: LU; chained equipercentile: CE; frequency estimation equipercentile: FE

Figure 20. Biology equating-difference plots (fall-fall random and fall-spring random or matched groups equatings minus fall-fall random groups Tucker criterion equating).

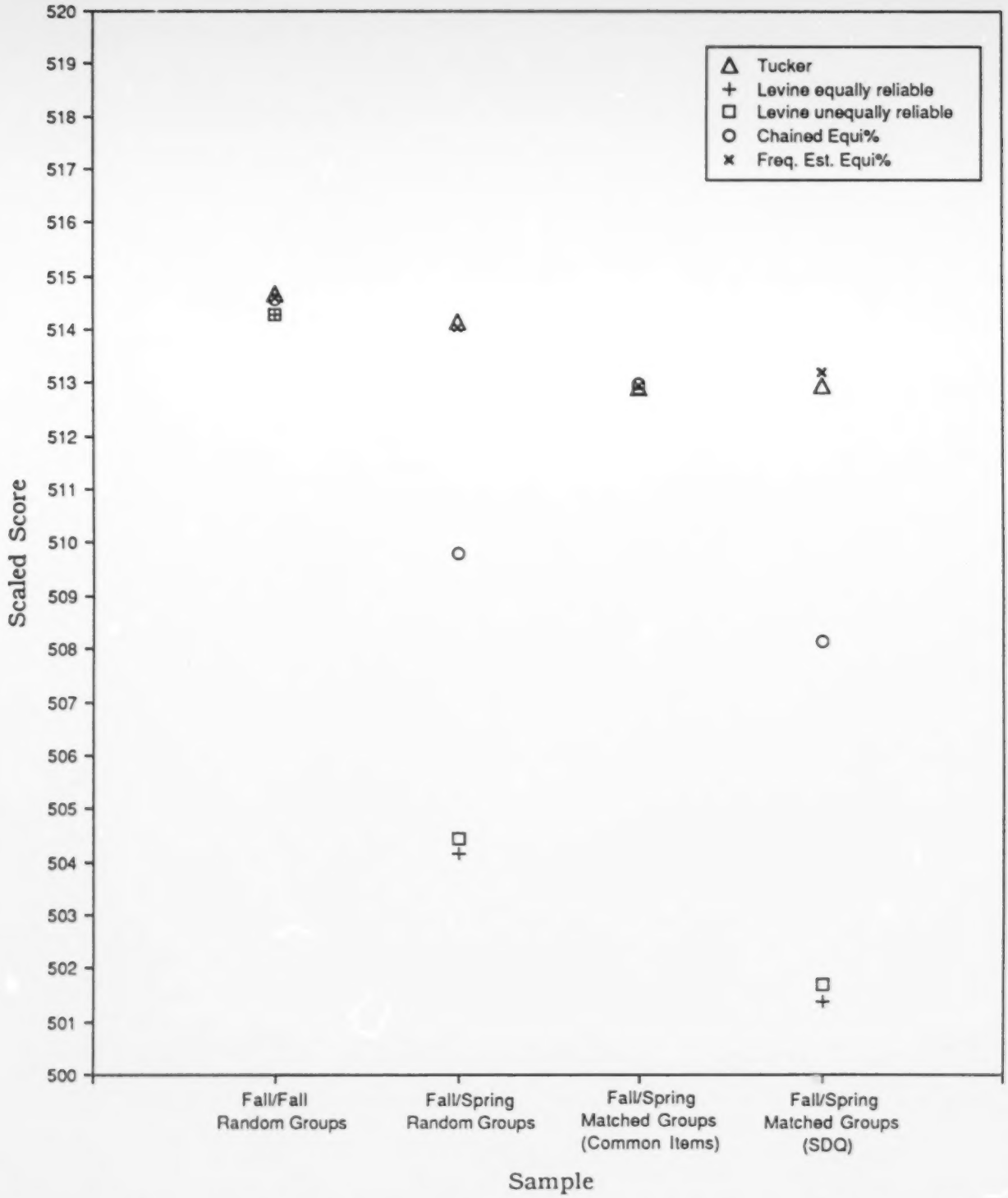
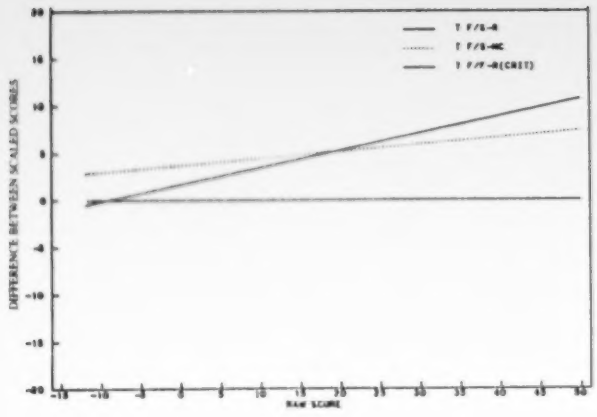
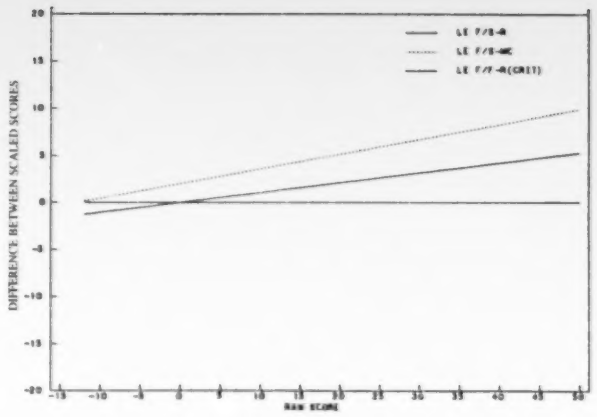


Figure 21. Plot of Biology projected new-form scaled-score means for all equating-method and equating-sample combinations.



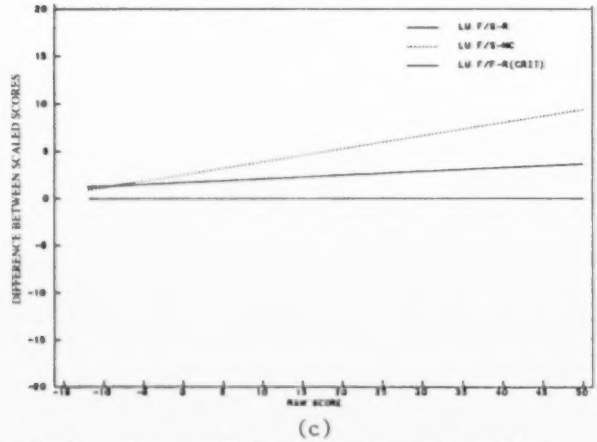
(a)

fall-spring random or matched groups Tucker equatings—fall-fall random groups Tucker criterion equating



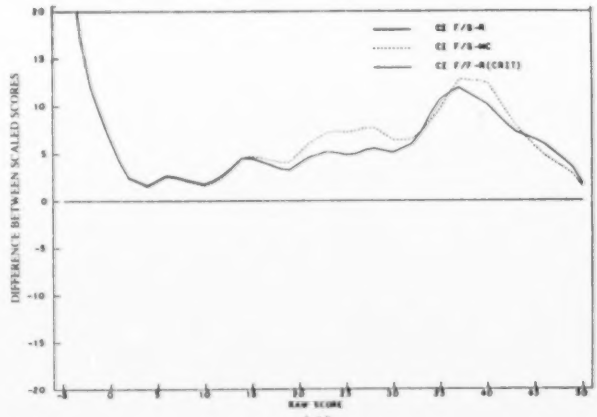
(b)

fall-spring random or matched groups Levine equally reliable equatings—fall-fall random groups Levine equally reliable criterion equating



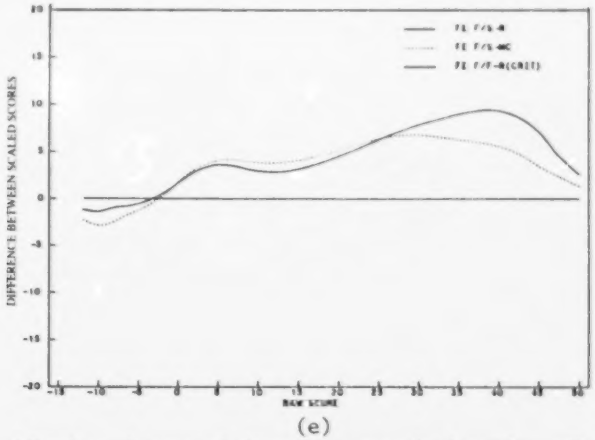
(c)

fall-spring random or matched groups Levine unequally reliable equatings—fall-fall random groups Levine unequally reliable criterion equating



(d)

fall-spring random or matched groups chained equipercentile equatings—fall-fall random groups chained equipercentile criterion equating

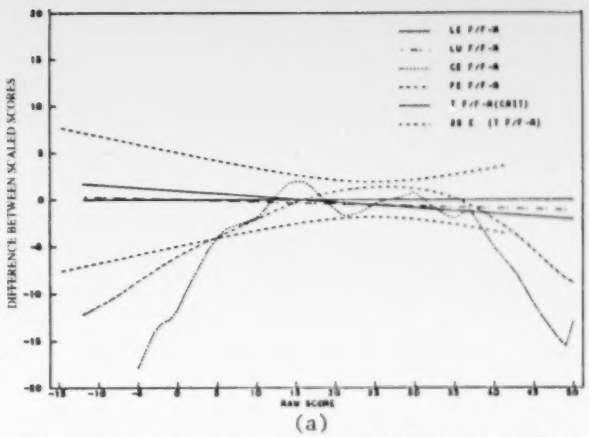


(e)

fall-spring random or matched groups frequency estimation equipercentile equatings—fall-fall random groups frequency estimation equipercentile criterion equating

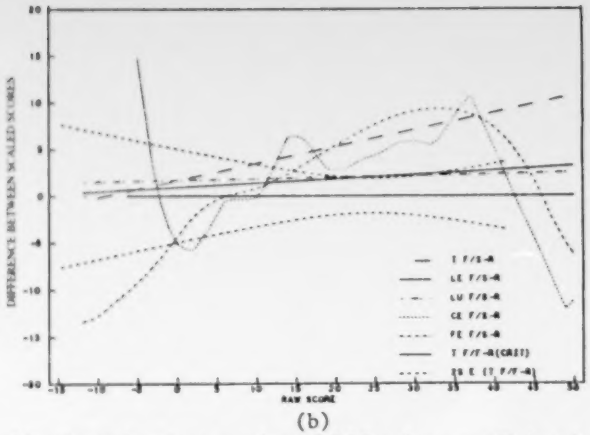
Abbreviations used in plots: fall-fall random groups: F/F-R; fall-spring random groups: F/S-R; fall-spring matched groups (common items): F/S-MC; Tucker: T; Levine equally reliable: LE; Levine unequally reliable: LU; chained equipercentile: CE; frequency estimation equipercentile: FE

Figure 22. Mathematics Level II equating-difference plots (fall-spring random or matched groups equatings minus fall-fall random groups criterion equating).



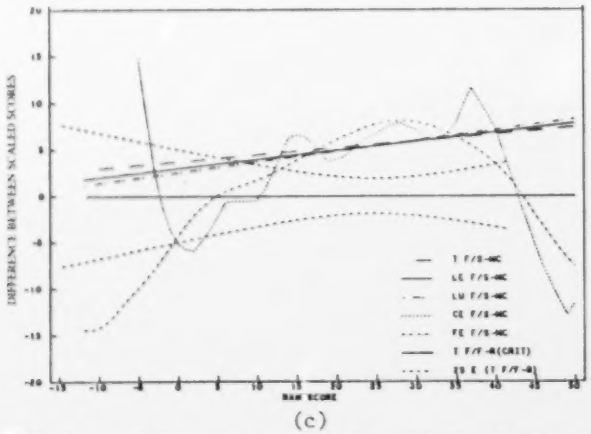
(a)

other fall-fall random groups equatings—fall-fall random groups Tucker criterion equating



(b)

fall-spring random groups equatings—fall-fall random groups Tucker criterion equating



(c)

fall-spring matched (common items) equatings—fall-fall random groups Tucker criterion equating

Abbreviations used in plots: fall-fall random groups: F/F-R; fall-spring random groups: F/S-R; fall-spring matched groups (common items): F/S-MC; Tucker: T; Levine equally reliable: LE; Levine unequally reliable: LU; chained equipercentile: CE; frequency estimation equipercentile: FE

Figure 23. Mathematics Level II equating-difference plots (fall-fall random and fall-spring random or matched groups equatings minus fall-fall random groups Tucker criterion equating).

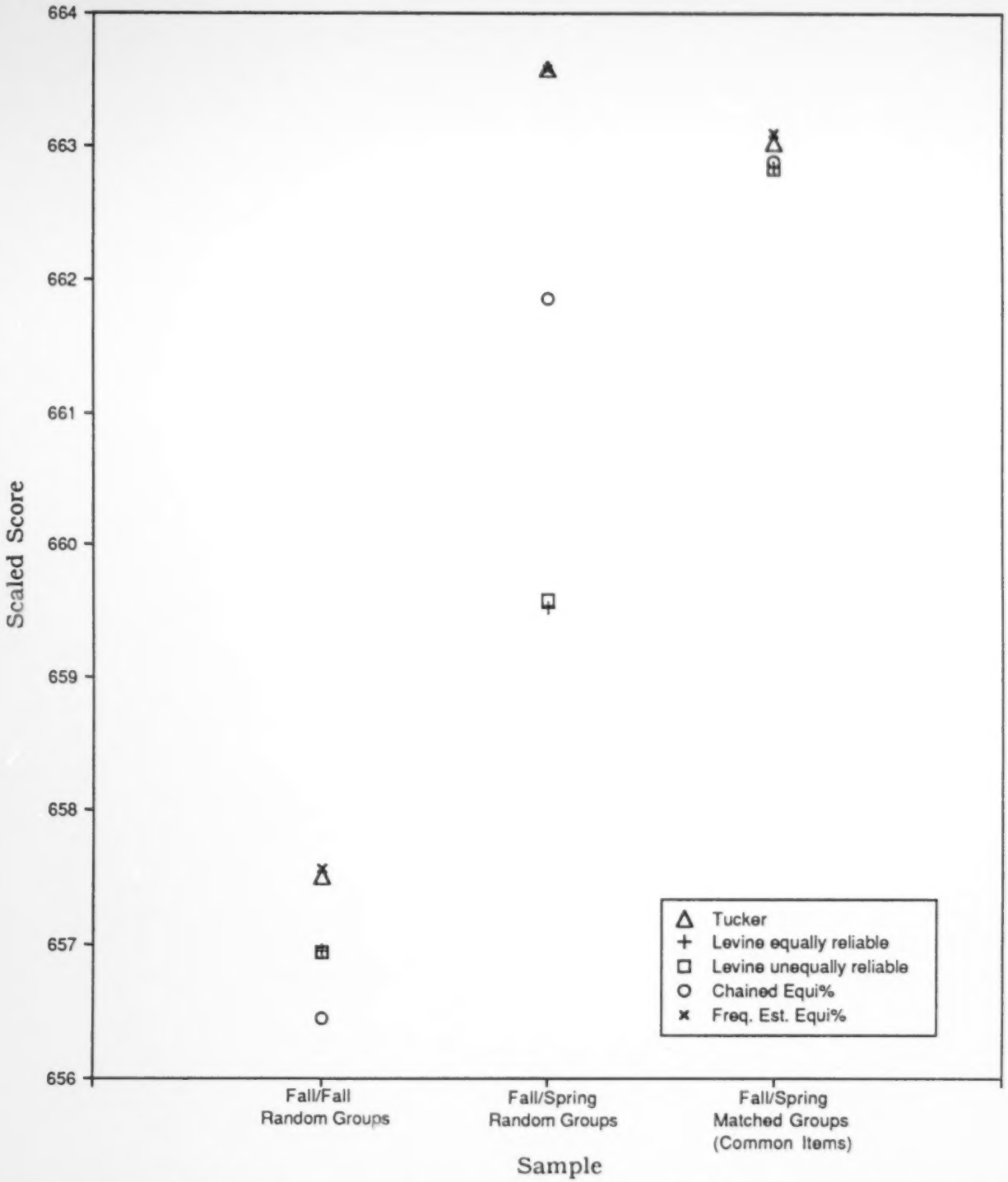
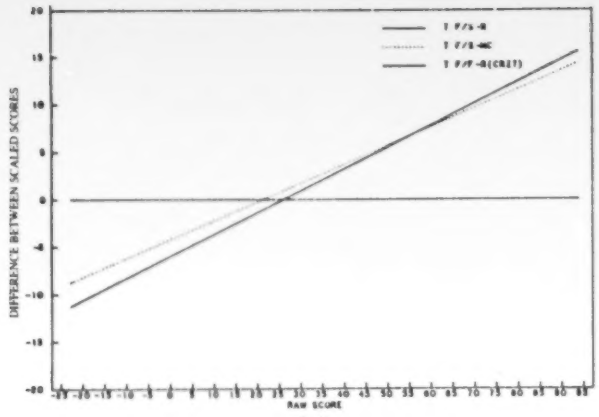
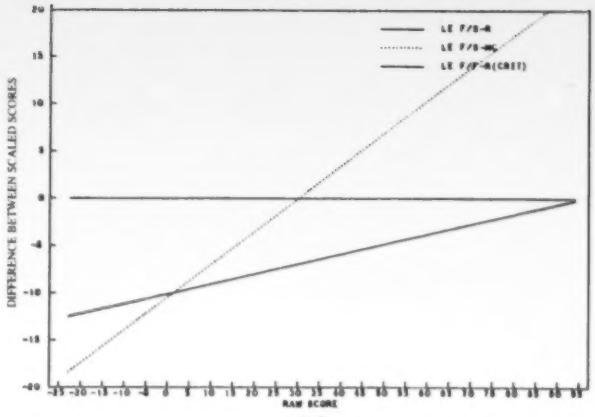


Figure 24. Plot of Mathematics Level II projected new-form scaled-score means for all equating-method and equating-sample combinations.



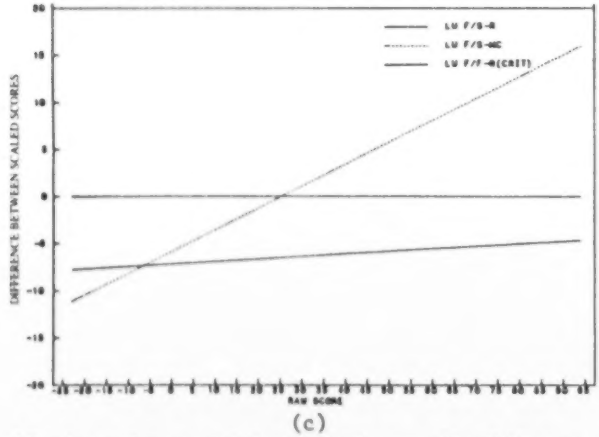
(a)

fall-spring random or matched groups Tucker equatings—fall-fall random groups Tucker criterion equating



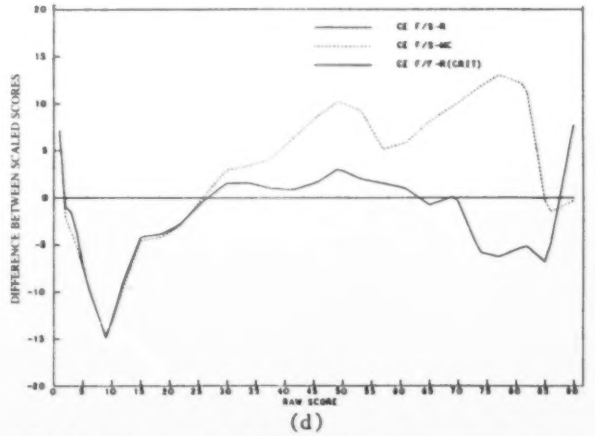
(b)

fall-spring random or matched groups Levine equally reliable equatings—fall-fall random groups Levine equally reliable criterion equating



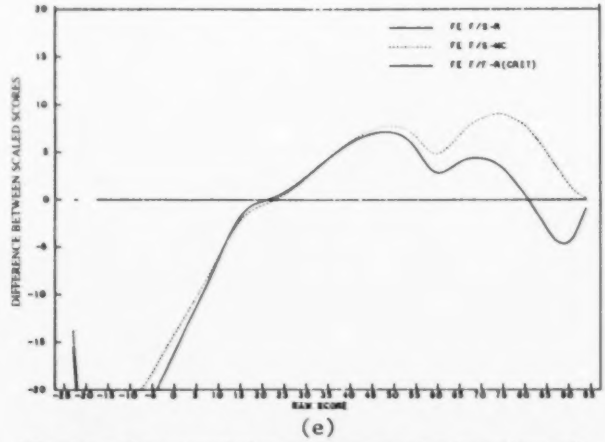
(c)

fall-spring random or matched groups Levine unequally reliable equatings—fall-fall random groups Levine unequally reliable criterion equating



(d)

fall-spring random or matched groups chained equipercentile equatings—fall-fall random groups chained equipercentile criterion equating

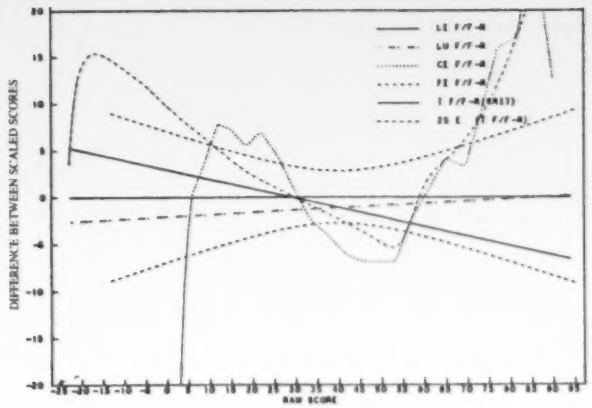


(e)

fall-spring random or matched groups frequency estimation equipercentile equatings—fall-fall random groups frequency estimation equipercentile criterion equating

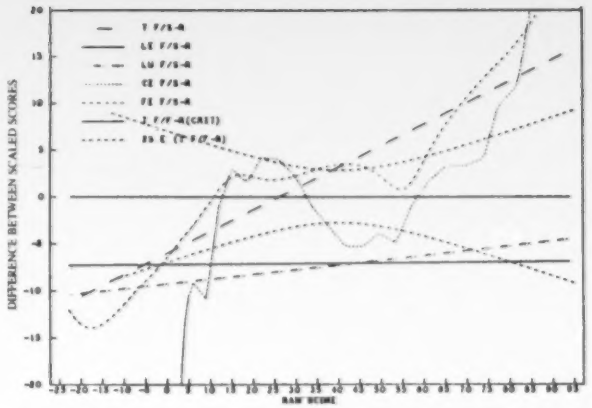
Abbreviations used in plots: fall-fall random groups: F/F-R; fall-spring random groups: F/S-R; fall-spring matched groups (common items): F/S-MC; Tucker: T; Levine equally reliable: LE; Levine unequally reliable: LU; chained equipercentile: CE; frequency estimation equipercentile: FE

Figure 25. American History and Social Studies equating-difference plots (fall-spring random or matched groups equatings minus fall-fall random groups criterion equating).



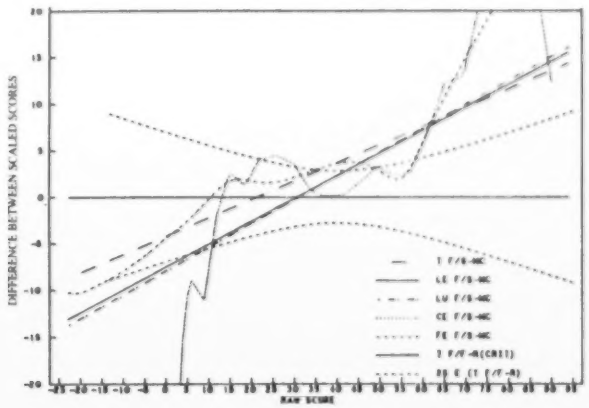
(a)

other fall-fall random groups equatings—fall-fall random groups Tucker criterion equating



(b)

fall-spring random groups equatings—fall-fall random groups Tucker criterion equating



(c)

fall-spring matched (common items) equatings—fall-fall random groups Tucker criterion equating

Abbreviations used in plots: fall-fall random groups: F/F-R; fall-spring random groups: F/S-R; fall-spring matched groups (common items): F/S-MC; Tucker: T; Levine equally reliable: LE; Levine unequally reliable: LU; chained equipercentile: CE; frequency estimation equipercentile: FE

Figure 26. American History and Social Studies equating-difference plots (fall-fall random and fall-spring random or matched groups equatings minus fall-fall random groups Tucker criterion equating).

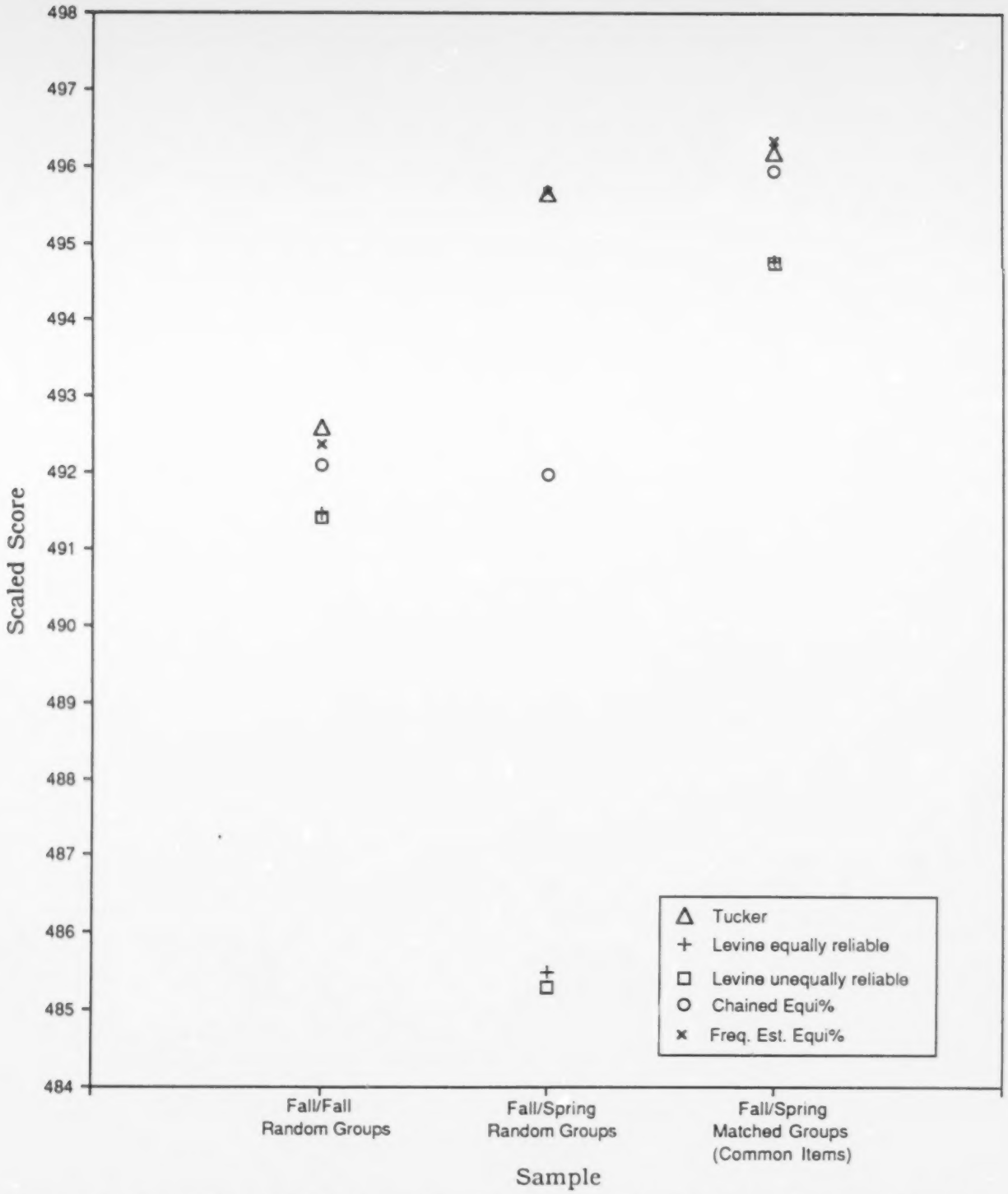
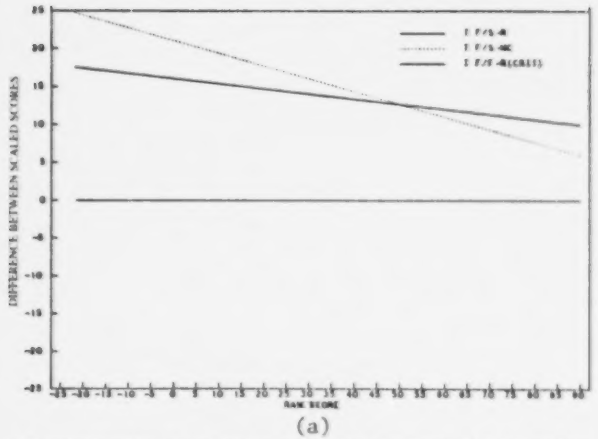
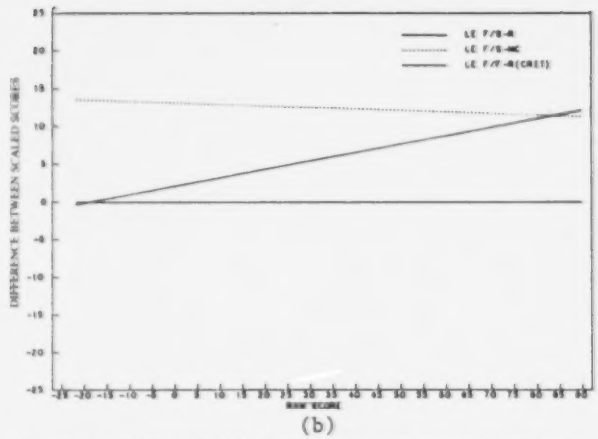


Figure 27. Plot of American History and Social Studies projected new-form scaled-score means for all equating-method and equating-sample combinations.



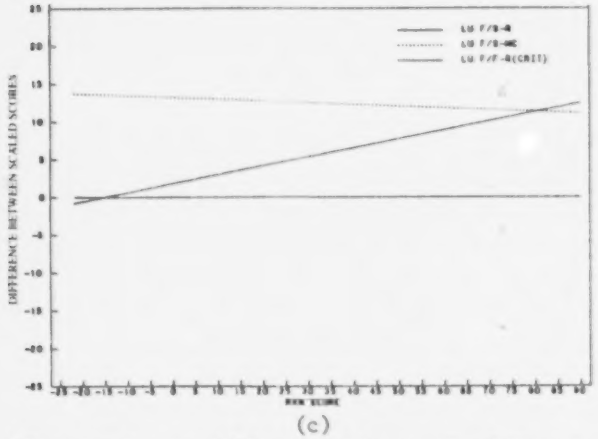
(a)

fall-spring random or matched groups Tucker equatings—fall-fall random groups Tucker criterion equating



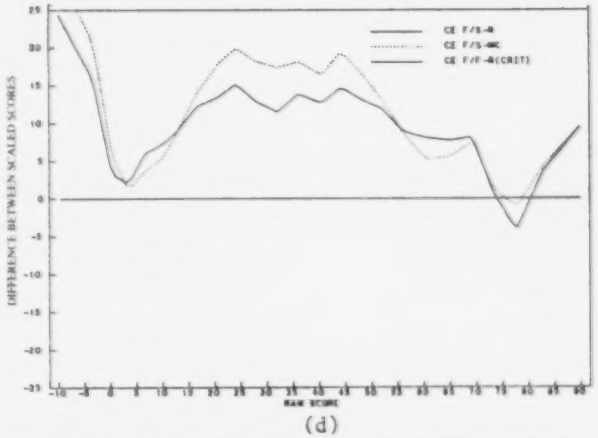
(b)

fall-spring random or matched groups Levine equally reliable equatings—fall-fall random groups Levine equally reliable criterion equating



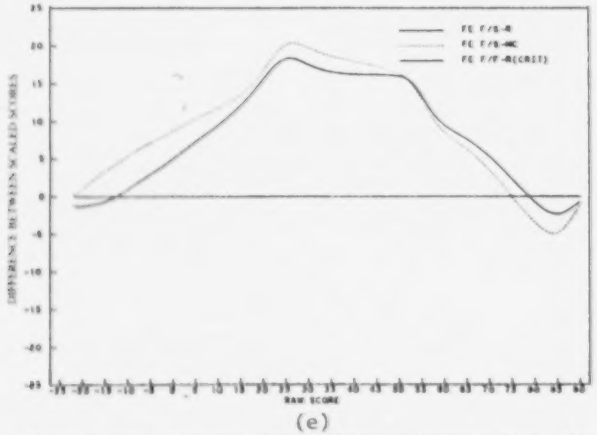
(c)

fall-spring random or matched groups Levine unequally reliable equatings—fall-fall random groups Levine unequally reliable criterion equating



(d)

fall-spring random or matched groups chained equipercentile equatings—fall-fall random groups chained equipercentile criterion equating

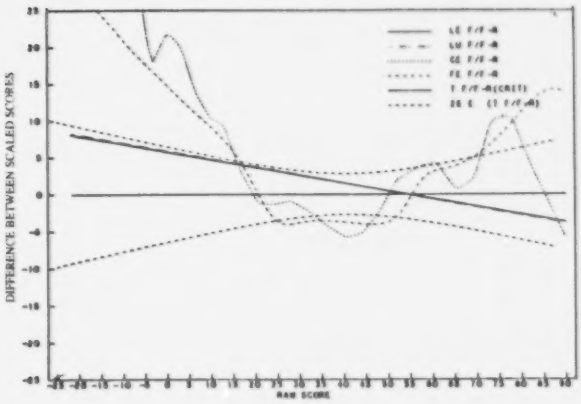


(e)

fall-spring random or matched groups frequency estimation equipercentile equatings—fall-fall random groups frequency estimation equipercentile criterion equating

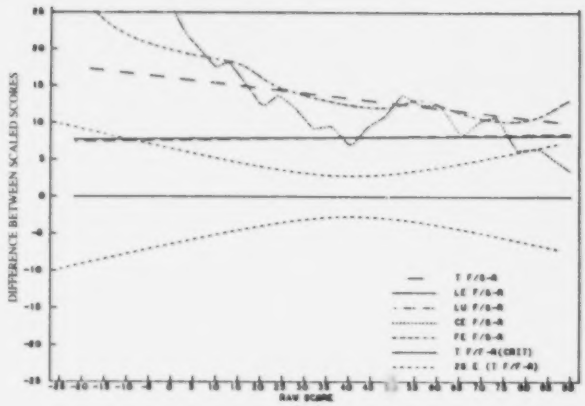
Abbreviations used in plots: fall-fall random groups: F/F-R; fall-spring random groups: F/S-R; fall-spring matched groups (common items): F/S-MC; Tucker: T; Levine equally reliable: LE; Levine unequally reliable: LU; chained equipercentile: CE; frequency estimation equipercentile: FE

Figure 28. Chemistry equating-difference plots (fall-spring random or matched groups equatings minus fall-fall random groups criterion equating).



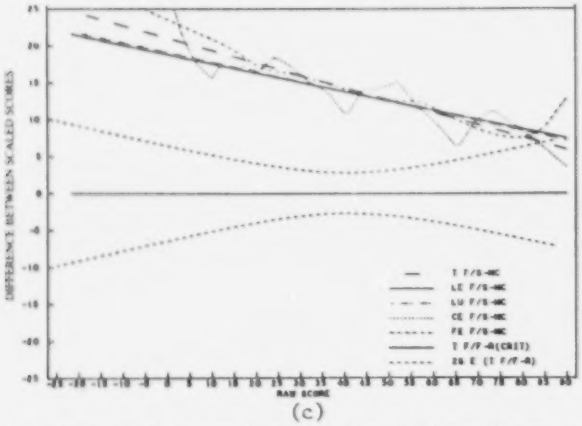
(a)

other fall-fall random groups equatings—fall-fall random groups Tucker criterion equating



(b)

fall-spring random groups equatings—fall-fall random groups Tucker criterion equating



(c)

fall-spring matched (common items) equatings—fall-fall random groups Tucker criterion equating

Abbreviations used in plots: fall-fall random groups: F/F-R; fall-spring random groups: F/S-R; fall-spring matched groups (common items): F/S-MC; Tucker: T; Levine equally reliable: LE; Levine unequallty reliable: LU; chained equipercentile: CE; frequency estimation equipercentile: FE

Figure 29. Chemistry equating-difference plots (fall-fall random and fall-spring random or matched groups equatings minus fall-fall random groups Tucker criterion equating).

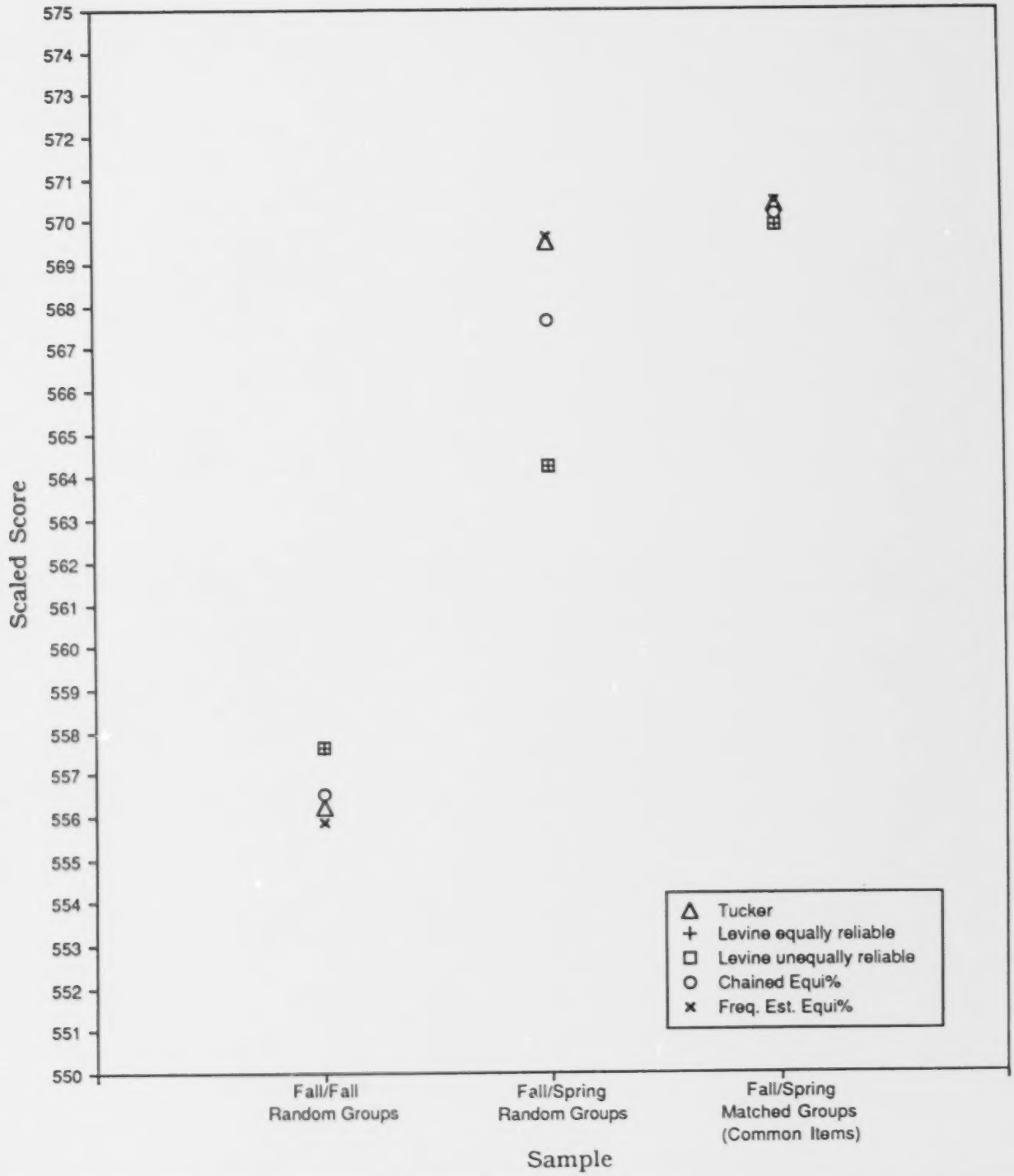
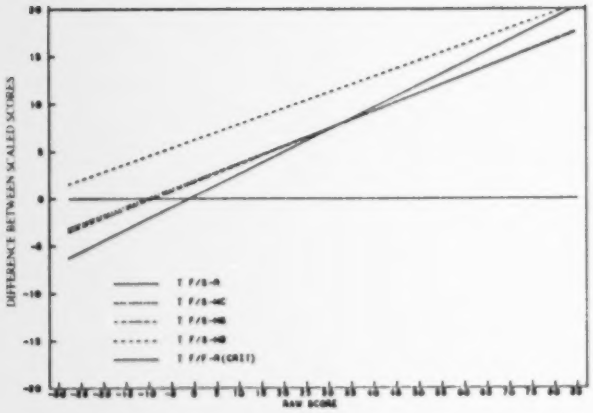
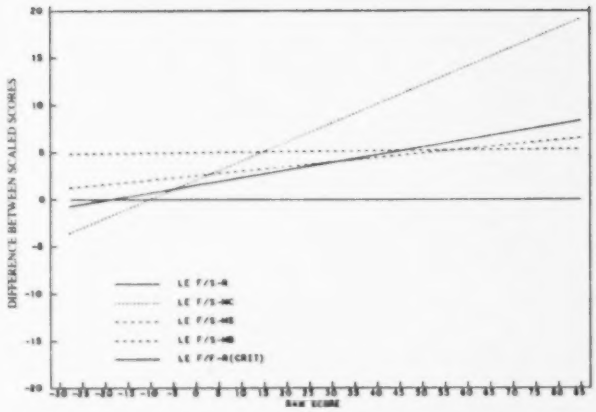


Figure 30. Plot of Chemistry projected new-form scaled-score means for all equating-method and equating-sample combinations.



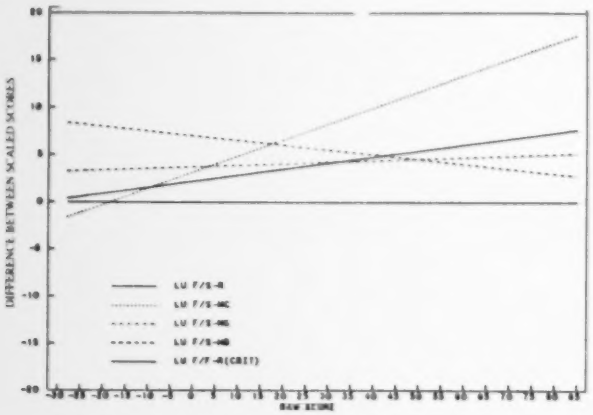
(a)

fall-spring random or matched groups Tucker equatings—fall-fall random groups Tucker criterion equating



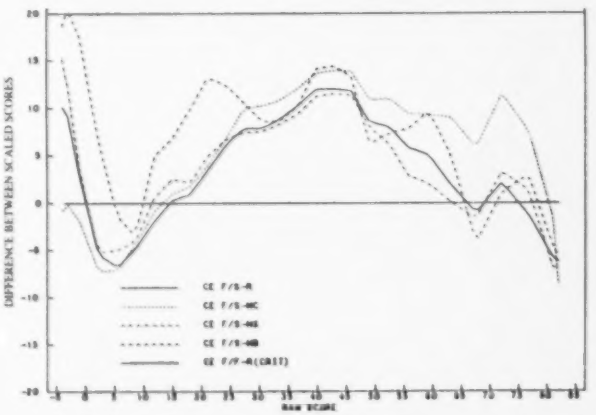
(b)

fall-spring random or matched groups Levine equally reliable equatings—fall-fall random groups Levine equally reliable criterion equating



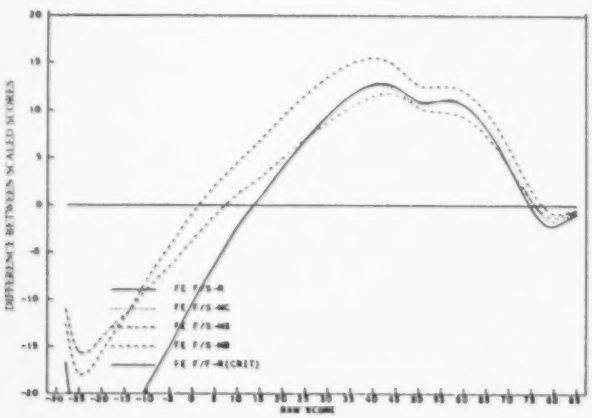
(c)

fall-spring random or matched groups Levine unequally reliable equating—fall-fall random groups Levine unequally reliable criterion equating



(d)

fall-spring random or matched groups chained equipercentile equatings—fall/fall random groups chained equipercentile criterion equating

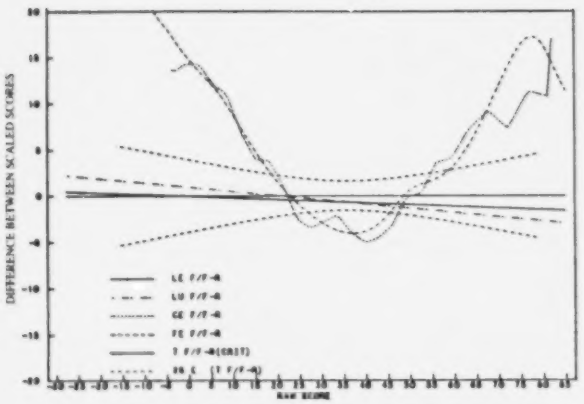


(e)

fall-spring random or matched groups frequency estimation equipercentile equatings—fall-fall random groups frequency estimation equipercentile criterion equating

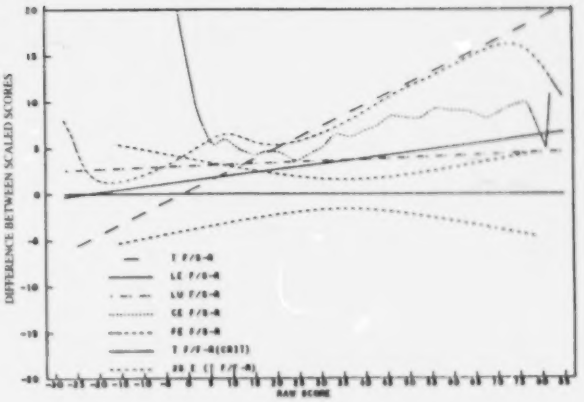
Abbreviations used in plots: fall-fall random groups: F/F-R; fall-spring random groups: F/S-R; fall-spring matched groups (common items): F/S-MC; fall-spring matched groups (SDQ): F/S-MS; fall-spring matched groups (BQ): F/S-MB; Tucker: T; Levine equally reliable: LE; Levine unequally reliable: LU; chained equipercentile: CE; frequency estimation equipercentile: FE

Figure 31. French equating-difference plots (fall-spring random or matched groups equatings minus fall-fall random groups criterion equating).



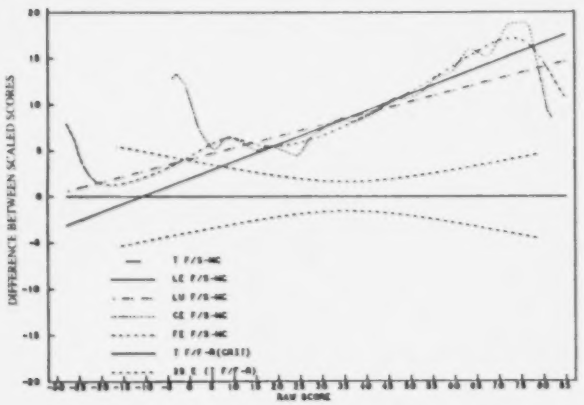
(a)

other fall-fall random groups equatings—fall-fall random groups Tucker criterion equating



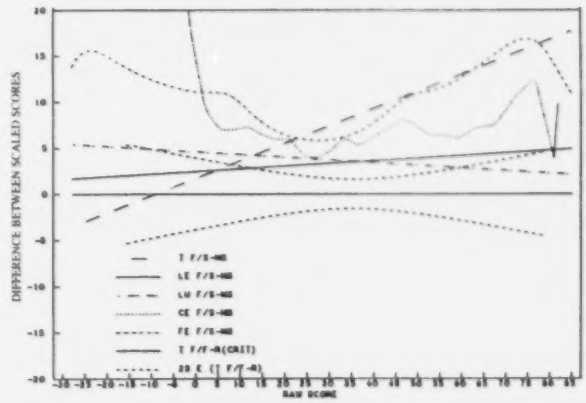
(b)

fall-spring random groups equatings—fall-fall random groups Tucker criterion equating



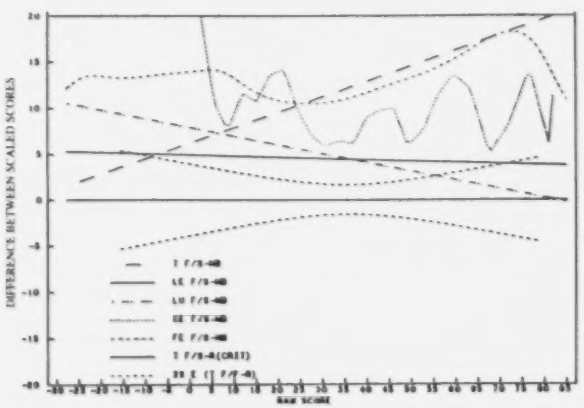
(c)

fall-spring matched (common items) equatings—fall-fall random groups Tucker criterion equating



(d)

fall-spring matched (SDQ) equatings—fall-fall random groups Tucker criterion equating



(e)

fall-spring matched (BQ) equatings—fall-fall random groups Tucker criterion equating

Abbreviations used in plots: fall-fall random groups: F/F-R; fall-spring random groups: F/S-R; fall-spring matched groups (common items): F/S-MC; fall-spring matched groups (SDQ): F/S-MS; fall-spring matched groups (BQ): F/S-MB; Tucker: T; Levine equally reliable: LE; Levine unequally reliable: LU; chained equipercentile: CE; frequency estimation equipercentile: FE

Figure 32. French equating-difference plots (fall-fall random and fall-spring random or matched groups equatings minus fall-fall random groups Tucker criterion equating).

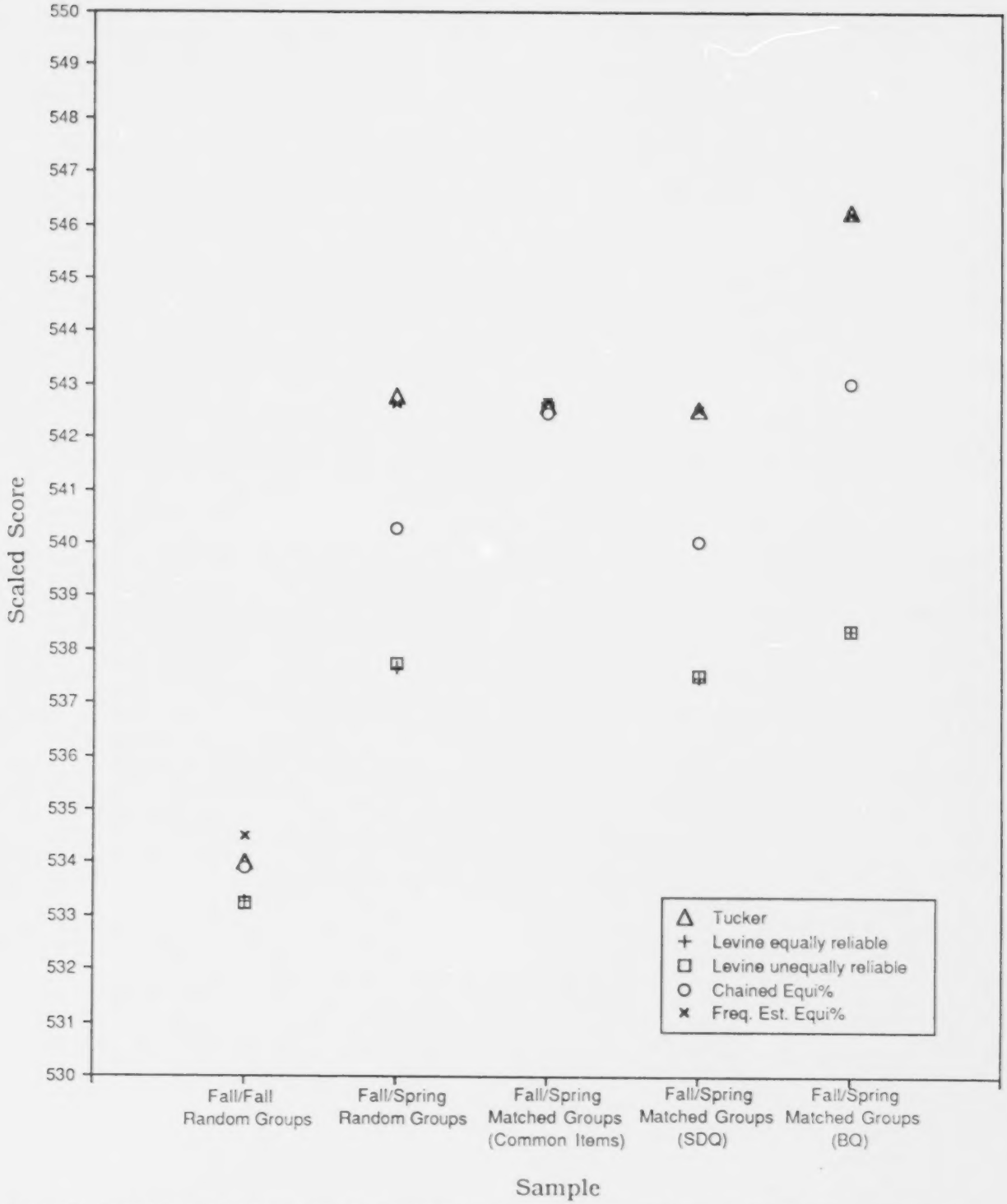


Figure 33. Plot of French projected new-form scaled-score means for all equating-method and equating-sample combinations.

Table 1. Numbers of Items in New and Old Forms and in Common-Item Sets for Tests in Study

Test	New Form		Old Form		Common-Item Set
	No. of Items	*Administered*	*No. of Items*	*Administered*	*No. of Items*
Biology	99	fall	95	fall, spring	36
Mathematics Level II	50	fall	50	fall, spring	19
American History and Social Studies	94	fall	100	fall, spring	20
Chemistry	90	fall	90	fall, spring	20
French	85	fall	85	fall, spring	21

Table 2a. Content and Skills Specifications for ATP Achievement Tests—Biology

Topics Covered	*Approximate Percentage of Test*
Cellular and Molecular Biology Cell structure and organization, mitosis, photosynthesis, cellular respiration, enzymes, molecular genetics, biosynthesis, biological chemistry	30
Ecology Energy flow, nutrient cycles, populations, communities, ecosystems, biomes	15
Classical Genetics Meiosis, Mendelian genetics, inheritance patterns	10
Organismal Structure and Function Anatomy and physiology, developmental biology, behavior	30
Evolution and Diversity Origin of life, evidence of evolution, natural selection, speciation, patterns of evolution, classification and diversity of prokaryotes, protists, fungi, plants, and animals	15

Skills Specifications		*Approximate Percentage of Test*
Level I	Essentially Recall: remembering specific facts; demonstrating straightforward knowledge of information and familiarity with terminology	50
Level II	Essentially Application: understanding concepts and reformulating information into other equivalent terms; applying knowledge to unfamiliar and/or practical situations; solving problems using mathematical relationships	30
Level III	Essentially Interpretation: inferring and deducing from qualitative and quantitative data and integrating information to form conclusions; recognizing unstated assumptions	20

Table 2b. Content of ATP Achievement Tests—Mathematics Level II

Topics Covered	*Approximate Percentage of Test*
Algebra	18
Geometry	
Solid geometry	8
Coordinate geometry	12
Trigonometry	20
Functions	24
Miscellaneous	18

Table 2c. Content of ATP Achievement Tests—American History and Social Studies

Material Covered	*Approximate Percentage of Test*
Political History	30–34
Economic History	16–18
Social History	16–20
Intellectual and Cultural History	8–10
Foreign Policy	11–15
Social Science Concepts, Methods, and Generalizations	10–12

Periods Covered	*Approximate Percentage of Test*
Pre-Columbian History to 1789	18
1790 to 1898	35
1899 to the Present	35
Nonchronological	12

BEST COPY AVAILABLE

Table 2d. Content and Skills Specifications for ATP Achievement Tests—Chemistry

Topics Covered	*Approximate Percentage of Test*
I. Atomic Theory and Structure Periodic relationships	10
II. Nuclear Reactions	2
III. Chemical Bonding and Molecular Structure	11
IV. States of Matter and Kinetic Molecular Theory	9
V. Solutions Concentration units, solubility, and colligative properties	6
VI. Acids and Bases	9
VII. Oxidation-reduction and Electrochemistry	7
VIII. Stoichiometry Mole concept, Avogadro's number, empirical and molecular formulas, percentage composition, stoichiometric calculations, and limiting reagents	11
IX. Reaction Rates Rate equations and factors affecting rates	2
X. Equilibrium Mass action expressions, ionic equilibria, and LeChatelier's principle	6
XI. Thermodynamics Energy changes in chemical reactions, randomness, and criteria for spontaneity	4
XII. Descriptive Chemistry Physical and chemical properties of elements and their more familiar compounds, including simple examples from organic chemistry; periodic properties	16
XIII. Laboratory Equipment, procedures, observations, safety, calculations, and interpretation of results	7

Skills Specifications		*Approximate Percentage of Test*
Level I	Essentially Recall: remembering information and understanding facts	30
Level II	Essentially Application: applying knowledge to unfamiliar and/or practical situations; solving problems using mathematical relationships	55
Level III	Essentially Interpretation: inferring and deducing from qualitative and quantitative data and integrating information to form conclusions	15

Note: Every edition contains approximately five questions on equation balancing and/or predicting products of chemical reactions. These are distributed among the various content categories.

Table 2e. Skills Specifications for ATP Achievement Tests—French

Skills Specifications	*Approximate Percentage of Test*
Vocabulary in Context	30
Structure	40
Reading Comprehension	30

Table 3. Raw-Score Summary Statistics for Total Tests and the Common-Item Set for Samples Used in Equatings—Biology

		*Total Test**		*Common-Item Set†*			
Sample	N	*Mean*	*Standard Deviation*	*Mean*	*Mean as % of Maximum*	*Standard Deviation*	*Common Item/Total Test Correlation*
Fall new form (random)	2,408	46.33	18.26	17.52	48.7	7.52	.92
Fall old form (random)	2,554	44.97	18.11	17.71	49.2	7.60	.92
Spring old form (random)	2,504	54.15	17.55	21.85	60.7	7.23	.93
Spring old form (matched common items)	2,408	44.26	17.93	17.52	48.7	7.52	.93
Spring old form (matched SDQ)	1,348	54.17	17.76	22.14	61.5	7.32	.93

*The new-form total test consisted of 99 items, whereas the old-form test consisted of 95 items.
†The common-item set consisted of 36 items.

Table 4. Raw-Score Summary Statistics for Total Tests and the Common-Item Set for Samples Used in Equatings—Mathematics Level II

		*Total Test**		*Common-Item Set†*			
Sample	N	*Mean*	*Standard Deviation*	*Mean*	*Mean as % of Maximum*	*Standard Deviation*	*Common Item/Total Test Correlation*
Fall new form (random)	2,304	24.61	9.56	9.55	50.3	4.19	.92
Fall old form (random)	2,346	26.25	9.71	9.73	51.2	4.28	.92
Spring old form (random)	2,009	29.65	10.17	10.98	57.8	4.41	.93
Spring old form (matched common items)	1,443	26.63	9.56	9.60	50.5	4.15	.92

*The new- and old-form total tests consisted of 50 items.
†The common-item set consisted of 19 items.

Table 5. Raw-Score Summary Statistics for Total Tests and the Common-Item Set for Samples Used in Equatings—American History and Social Studies

		*Total Test**		*Common-Item Set†*			
Sample	N	*Mean*	*Standard Deviation*	*Mean*	*Mean as % of Maximum*	*Standard Deviation*	*Common Item/Total Test Correlation*
Fall new form (random)	2,078	39.64	16.14	8.49	42.5	4.15	.87
Fall old form (random)	2,102	40.30	16.60	8.67	43.4	4.28	.86
Spring old form (random)	2,031	46.93	17.92	10.38	51.9	4.48	.88
Spring old form (matched common items)	1,329	41.11	16.62	8.71	43.6	4.09	.86

*The new-form total test consisted of 94 items, whereas the old-form test consisted of 100 items.
†The common-item set consisted of 20 items.

Table 6. Raw-Score Summary Statistics for Total Tests and the Common-Item Set for Samples Used in Equatings—Chemistry

		*Total Test**		*Common-Item Set†*			
Sample	N	*Mean*	*Standard Deviation*	*Mean*	*Mean as % of Maximum*	*Standard Deviation*	*Common Item/Total Test Correlation*
Fall new form (random)	2,017	41.29	17.84	10.22	51.1	4.42	.88
Fall old form (random)	2,249	36.36	18.91	9.94	49.7	4.61	.89
Spring old form (random)	2,206	43.13	17.74	11.11	55.6	4.31	.87
Spring old form (matched common items)	1,436	40.43	17.54	10.31	51.6	4.36	.88

*The new- and old-form total tests consisted of 90 items.
†The common-item set consisted of 20 items.

Table 7. Raw-Score Summary Statistics for Total Tests and the Common-Item Set for Samples Used in Equatings—French

		*Total Test**		*Common-Item Set†*			
Sample	N	*Mean*	*Standard Deviation*	*Mean*	*Mean as % of Maximum*	*Standard Deviation*	*Common Item/Total Test Correlation*
Fall new form (random)	6,125	35.67	14.90	8.60	41.0	4.66	.87
Fall old form (random)	7,269	34.42	15.04	8.74	41.6	4.69	.88
Spring old form (random)	6,078	38.38	16.18	9.62	45.8	4.91	.89
Spring old form (matched common items)	3,248	35.39	15.42	8.60	41.0	4.66	.88
Spring old form (matched SDQ)	2,959	38.37	16.04	9.63	45.9	4.90	.89
Spring old form (matched BQ)	928	41.05	16.12	10.34	49.2	4.95	.90

*The new- and old-form total tests consisted of 85 items.
†The common-item set consisted of 21 items.

Table 8. Correlation Coefficients for Item-Difficulty Estimates (Deltas) and New-Form Total-Test Summary Statistics for Different Sample Combinations—Biology

		Total-Test Equated Delta†	
Sample Combination	r*	*Mean*	*S.D.*
Fall-fall random groups	.99	13.09	1.94
Fall-spring random groups	.73	13.09	1.94
Fall-spring matched groups (common items)	.79	13.08	1.94
Fall-spring matched groups (SDQ)	.77	13.09	1.94

*Correlation coefficients between the deltas obtained for the 36 common items when given to the new- and old-form samples indicated.
†Statistical specifications call for an equated delta mean of 13.0 and standard deviation of 2.2.

Table 9. Correlation Coefficients for Item-Difficulty Estimates (Deltas) and New-Form Total-Test Summary Statistics for Different Sample Combinations—Mathematics Level II

		Total-Test Equated Delta†	
Sample Combination	r*	*Mean*	*S.D.*
Fall-fall random groups	.99	15.52	2.47
Fall-spring random groups	.98	15.53	2.48
Fall-spring matched groups (common items)	.98	15.53	2.47

*Correlation coefficients between the deltas obtained for the 19 common items when given to the new- and old-form samples indicated.
†Statistical specifications call for an equated delta mean of 15.2 and standard deviation of 2.2.

Table 10. Correlation Coefficients for Item-Difficulty Estimates (Deltas) and New-Form Total-Test Summary Statistics for Different Sample Combinations—American History and Social Studies

		Total-Test Equated Delta†	
Sample Combination	r*	*Mean*	*S.D.*
Fall-fall random groups	.94	11.34	2.35
Fall-spring random groups	.93	11.34	2.34
Fall-spring matched groups (common items)	.93	11.34	2.36

*Correlation coefficients between the deltas obtained for the 20 common items when given to the new- and old-form samples indicated.
†Statistical specifications call for an equated delta mean of 11.5 and standard deviation of 2.2.

Table 11. Correlation Coefficients for Item-Difficulty Estimates (Deltas) and New-Form Total-Test Summary Statistics for Different Sample Combinations—Chemistry

		Total-Test Equated Delta†	
Sample Combination	r*	*Mean*	*S.D.*
Fall-fall random groups	.97	13.14	1.84
Fall-spring random groups	.94	13.15	1.82
Fall-spring matched groups (common items)	.94	13.16	1.85

*Correlation coefficients between the deltas obtained for the 20 common items when given to the new- and old-form samples indicated.
†Statistical specifications call for an equated delta mean of 13.0 and standard deviation of 2.0.

Table 12. Correlation Coefficients for Item-Difficulty Estimates (Deltas) and New-Form Total-Test Summary Statistics for Different Sample Combinations—French

		Total-Test Equated Delta†	
Sample Combination	r*	*Mean*	*S.D.*
Fall-fall random groups	.99	11.69	3.45
Fall-spring random groups	.99	11.68	3.45
Fall-spring matched groups (common items)	.99	11.69	3.44
Fall-spring matched groups (SDQ)	.99	11.69	3.46
Fall-spring matched groups (BQ)	.98	11.58	3.65

*Correlation coefficients between the deltas obtained for the 21 common items when given to the new- and old-form samples indicated.
†Statistical specifications call for an equated delta mean of 11.7 and standard deviation of 2.2.

Table 13. New-Form Scaled-Score Summary Statistics Resulting from Equating-Method and Equating-Sample Combinations—Biology

	Equating Method									
	Tucker		*Levine Equally Reliable*		*Levine Unequally Reliable*		*Chained Equipercentile*		*Frequency Estimation Equipercentile*	
Samples	*Mean*	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*
Fall-fall random groups	514.67	103.64	514.27	103.30	514.27	103.60	514.58	103.42	514.59	103.76
Fall-spring random groups	514.13	104.82	504.15	106.00	504.45	105.11	509.79	104.23	514.05	105.08
Fall-spring matched groups (common items)	512.91	103.46	512.91	103.46	512.92	102.97	512.97	103.31	512.95	103.46
Fall-spring matched groups (SDQ)	512.93	105.22	501.37	105.94	501.70	104.77	508.15	102.73	513.20	104.41

Raw-score frequency distributions used to compute scaled-score summary data were obtained from fall new-form total group (N = 7,208).

Table 14. New-Form Scaled-Score Summary Statistics Resulting from Equating-Method and Equating-Sample Combinations—Mathematics Level II

	Equating Method									
	Tucker		*Levine Equally Reliable*		*Levine Unequally Reliable*		*Chained Equipercentile*		*Frequency Estimation Equipercentile*	
Samples	*Mean*	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*
Fall-fall random groups	657.51	79.68	656.96	79.10	656.94	79.46	656.45	78.93	657.56	79.88
Fall-spring random groups	663.58	81.39	659.53	80.10	659.58	79.83	661.86	80.73	663.57	81.86
Fall-spring matched groups (common items)	663.02	80.37	662.84	80.59	662.83	80.75	662.88	80.91	663.10	80.58

Raw-score frequency distributions used to compute scaled-score summary data were obtained from fall new-form total group (N = 13,825).

Table 15. New-Form Scaled-Score Summary Statistics Resulting from Equating-Method and Equating-Sample Combinations—American History and Social Studies

	Equating Method									
	Tucker		*Levine Equally Reliable*		*Levine Unequally Reliable*		*Chained Equipercentile*		*Frequency Estimation Equipercentile*	
Samples	*Mean*	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*
Fall-fall random groups	492.58	89.07	491.47	87.42	491.42	89.46	492.10	87.92	492.38	88.83
Fall-spring random groups	495.66	92.81	485.50	89.13	485.30	89.90	491.99	89.14	495.68	90.93
Fall-spring matched groups (common items)	496.18	92.28	494.76	93.05	494.75	93.23	495.95	92.42	496.34	92.04

Raw-score frequency distributions used to compute scaled-score summary data were obtained from fall new-form total group (N = 16,598).

Table 16. New-Form Scaled-Score Summary Statistics Resulting from Equating-Method and Equating-Sample Combinations—Chemistry

	Equating Method									
	Tucker		*Levine Equally Reliable*		*Levine Unequally Reliable*		*Chained Equipercentile*		*Frequency Estimation Equipercentile*	
Samples	*Mean*	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*
Fall-fall random groups	556.21	94.25	557.63	92.36	557.63	92.31	556.54	94.37	555.87	94.84
Fall-spring random groups	569.48	93.04	564.25	94.36	564.24	94.44	567.65	92.60	569.63	92.37
Fall-spring matched groups (common items)	570.39	91.25	569.88	92.00	569.88	91.89	570.13	91.49	570.45	91.10

Raw-score frequency distributions used to compute scaled-score summary data were obtained from fall new-form total group (N = 8,059).

Table 17. New-Form Scaled-Score Summary Statistics Resulting from Equating-Method and Equating-Sample Combinations—French

	Equating Method									
	Tucker		*Levine Equally Reliable*		*Levine Unequally Reliable*		*Chained Equipercentile*		*Frequency Estimation Equipercentile*	
Samples	*Mean*	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*
Fall-fall random groups	534.00	103.59	533.26	103.29	533.25	102.84	533.92	103.22	534.52	104.05
Fall-spring random groups	542.78	107.44	537.68	104.62	537.73	103.89	540.27	104.45	542.64	106.50
Fall-spring matched groups (common items)	542.62	106.60	542.61	106.60	542.58	105.64	542.45	105.99	542.67	106.57
Fall-spring matched groups (SDQ)	542.54	106.68	537.50	104.06	537.52	103.11	540.04	103.57	542.54	105.52
Fall-spring matched groups (BQ)	546.28	106.34	538.37	103.37	538.38	102.02	543.04	102.43	546.22	104.84

Raw-score frequency distributions used to compute scaled-score summary data were obtained from fall new-form total group (N = 7,310).

BEST COPY AVAILABLE